Bayesian Statistics

Dr Nic Freeman

October 29, 2024

©Nic Freeman, University of Sheffield, 2024.

©Nic Freeman, University of Sheffield, 2024.

Contents

©Nic Freeman, University of Sheffield, 2024.

©Nic Freeman, University of Sheffield, 2024.

Chapter 0

Introduction

0.1 Organization

0.1.1 Syllabus

These notes are for two courses: MAS364 and MAS61006. All of the material in these notes is shared between both courses, but only MAS61006 students continue with the course next semester. Students on MAS61006 have some extra computer sessions this semester in preparation for that, and will do their project work next semester (and not this semester), but otherwise the teaching this semester is identical.

Some parts of the notes are marked with a (⊘) symbol, which means they are off-syllabus, for everyone. These cover some tangential material, technical proofs and other optional content that is included purely for interest.

0.1.2 End-of-chapter questions and problem sheets

At the end of chapter of these notes there is a set of problems for you to solve. The questions are marked with stars to indicate their rough level of difficulty:

⋆ I think (or rather, hope) that many people will find this question easy.

⋆⋆ I think this question is within the usual range of difficulty.

⋆⋆⋆ For whatever reason, I think many people will find this question difficult.

Most of the questions have two stars.

All of the solutions to the end-of-chapter questions within these notes are provided at the end of the online version of the notes (and not in the paper version) in Appendix C. You should work through these questions as we go through the chapters, and review your own solutions using the typed solutions.

At two points during the semester, a problem sheet of additional exercises will be set. About one week later, we will go through the questions in lectures, and you should self-mark your solutions. These do not count towards your final mark.

©Nic Freeman, University of Sheffield, 2024.

0.1.3 Assessment

- Students taking MAS364: you will have a two hour exam in January. You also have a project, involving R or Python (you may choose which), towards the end of the autumn semester. Details of the project will be released part-way through the autumn. The exam counts for 85% of your final mark, and the project for 15%.
- Students taking MAS61006: you will have a three hour exam in the summer, which will also include questions from the second semester. You will do project work as part of the second semester and *you do not do the project this semester*. The exam counts for 60% of your final mark, and the project work in the second semester counts for 40%.

In both cases, all questions on the exams will be compulsory. Some advice on how to structure your revision can be found in Appendix B of these notes.

0.1.4 Website

Further information, including the timetable, can be found on

`https://nicfreeman1209.github.io/Website/MASx64/`.

©Nic Freeman, University of Sheffield, 2024.

0.2 Outline of the course

Bayesian learning is the process of using data to update statistical models. The key principle is that

$$(\text{a model})|_{\{\text{model = the data we observed}\}} \stackrel{\text{d}}{=} (\text{a better model}). \tag{0.1}$$

where $|_{\{...\}}$ denotes conditioning, in the sense of conditional probability. The process of finding the right hand side, given all the inputs on the left hand side, is known as a Bayesian update. Performing one or more such updates in succession is known as Bayesian learning.

We begin our course with an introduction to conditional probability in Chapter 1. We introduce Bayesian statistical models in Chapters 2 (for discrete data) and 3 (for continuous data). These models are similar to those that you will already be familiar with, with the modification that we treat the parameters of the model as random variables. The operation in (0.1) acts to 'update' these parameters, to make the model better fit the data. In Chapter 8 we introduce a computational framework in which the operation (0.1) can be computed numerically, in full generality.

Bayesian learning is the oldest form of statistical learning and is often traced back to work of Laplace (1749-1827), but its modern treatment is very different to its history. Before the advent of modern computers the general methods in Chapter 8 were not available, and it was (consequently) not possible to perform Bayesian updates except in simple situations. This led to a period of several decades where statisticians developed approximation theorems, and that theory gave birth to most of the non-Bayesian statistical methods that are still widely used today – for example, maximum likelihood estimators, p-values, confidence intervals, t-tests and so on. Because these methods depend on approximation theorems, their results can be hard to interpret and their accuracy depends upon complicated conditions that are difficult to check. For example, it is very common to see p-values and confidence intervals misinterpreted, or to see misunderstandings of the output of well-known statistical tests.

The Bayesian framework avoids most of these difficulties by working directly with conditional probabilities. It has been growing in popularity ever since computers became widely available and may, in time, supplant older methods entirely. The only trade off is that, except for some special cases, it requires complex numerical methods to implement.

Returning to our own course; in Chapter 4 we study the cases in which Bayesian updates can be performed without the aid of computers. We will use such cases mainly as a way to better understand how Bayesian models behave. We then study the choice of 'prior' in Chapter 5. This provides a framework for incorporating pre-existing beliefs, known as priors, into statistical analysis. These beliefs may come in a convenient mathematical form, or may need to be elicited from subject experts with (perhaps) little understanding of statistics. We study the related framework of statistical testing in Chapter 7.

We discuss the relationship between Bayesian inference and other statistical methods in Chapter 6. Broadly, we build up a picture which shows that many branches of statistics can be viewed as simplifications of Bayesian methods. In that sense, Bayesian methods are the most natural form of statistical inference.

©Nic Freeman, University of Sheffield, 2024.

Chapter 1

Conditioning

1.1 Random variables

Let X be a random variable taking values in $\mathbb{R}$. You should think of X as an object that takes a random value, which is hopefully natural. Most of the things we interact with are random e.g. when we buy a pair of shoes we do not know how long they will last for; when we walk home later, we do not know how much rain will fall, and so on. In principle we might think of anything as being random, but within this course we will restrict ourselves to random variables that take values in $\mathbb{R}^d$. We won't use bold symbols for vectors in this course. Typically we will write x or y for elements of $\mathbb{R}^d$, and when we need to use coordinates we'll write e.g. $x = (x_1, \dots x_d) \in \mathbb{R}^d$, where $x_i \in \mathbb{R}$.

We are interested in two particular types of random variable in this course, captured by the following definition.

Definition 1.1.1 Let X be a random variable taking values in $\mathbb{R}^d$.

1. We say that X is *discrete* if there exists a countable set $A \subseteq \mathbb{R}^d$ such that $\mathbb{P}[X \in A] = 1$.

   In the case $d = 1$, this will usually mean that either $P[X \in \mathbb{N}]$ or $\mathbb{P}[X \in \mathbb{Z}] = 1$. We use the terminology 'let X be random variable with values in $\mathbb{N}$ (or $\mathbb{Z}$)' for this case.

   In this case the function $p_X(x) = \mathbb{P}[X = x]$, defined for $x \in \mathbb{R}^d$, is known as the *probability mass function* or simply p.m.f. of X.

   The *range* of X is the set $R_X = \{x \in \mathbb{R}^d \,;\, \mathbb{P}[X = x] > 0\}$.

2. We say that X is *continuous* if there exists a function $f_X : \mathbb{R}^d \to [0, \infty)$ such that

$$\mathbb{P}[X \in A] = \int_A f_X(x)\, dx \tag{1.1}$$

   for all $A \subseteq \mathbb{R}^d$.

   In this case f_X is known as the *probability density function* or simply p.d.f. of X. For $d > 1$ it is common to write $X = (X_1, \dots, X_d)$ and refer to $f_X(x)$ as the *joint* p.d.f. of the X_i.

   The *range* of X is the set $R_X = \{x \in \mathbb{R}^d \,;\, f_X(x) > 0\}$.

Most random variables used in statistical inference are one of these two types. In this course we will use reference sheets of named distributions, found in Appendix A, covering a very large

©Nic Freeman, University of Sheffield, 2024.

range of examples. These reference sheets will be made available in the exam. You should be familiar with relationships between named distributions that were discussed in earlier courses, for example the relationship between Bernoulli trials and the Geometric and Binomial distributions.

Note that the integral in (1.1) is over a set $A \subseteq \mathbb{R}^d$, with variable $x \in \mathbb{R}^d$. We'll generally use this notation instead of writing out multiple integral signs (e.g. $\int \int \int \cdots \int \ldots dx_1\, dx_2, \ldots, dx_d$) in this course.

Definition 1.1.2 Let X be a random variable taking values in $\mathbb{R}^d$. We say that a random variable X is *deterministic* if there exists $x \in \mathbb{R}^d$ such that $\mathbb{P}[X = x] = 1$.

We will often view a constant, say $a \in \mathbb{R}$, as an example of a deterministic random variable. This is another slight abuse of terminology, but it is natural and it won't cause any trouble. Note that deterministic random variables are a special type of discrete random variable.

©Nic Freeman, University of Sheffield, 2024.

1.1.1 (⊘) Technicalities

In this off-syllabus section we mention three technical points. They are aimed mainly at students with more technical backgrounds in analysis and probability theory. We won't discuss these points in lectures.

1. More advanced textbooks use the term *absolutely continuous* for the class of random variables that we have called continuous. The complication arises because there are random variables for which F_X is a continuous function but no p.d.f. f_X exists. These random variables are usually associated to random fractals and are rarely used within statistics, so in statistics it is common to drop the word 'absolutely'.

2. In this course we will use the convention that probability density functions must be continuous (as functions) except where they are zero. You can check that all of the distributions on the reference sheet in Appendix A are given in this form.

   In fact, probability density functions $f_X(x)$ are only defined *almost everywhere*. The term *for almost all* x is also commonly used. We cannot explain the precise meaning of it within this course, and many (otherwise good) textbooks on Bayesian statistics fail to note that this difficulty exists. Loosely, the same distribution can be defined using two (or more) different probability density functions $f_X(x)$ and $f'_X(x)$, but it will always be the case that $f_X(x) = f'_X(x)$ for 'almost all' values of x. We will discuss the matter further in Section 1.2, Remarks 3.1.3 and 6.1.2.

3. In our definitions and results above, the sets A for which we evaluate $\mathbb{P}[X \in A]$ must be Borel subsets of $\mathbb{R}^d$. In practice this technicality does not restrict us at all and we will continue to ignore this point for the remainder of the course.

Taking care of these issues rigorously requires some background on Lebesgue integration, but we do not assume that background for this course.

©Nic Freeman, University of Sheffield, 2024.

1.2 Equality in distribution

Let X be a random variable taking values in $\mathbb{R}^d$. The *law* or *distribution* of X is the function $A \mapsto \mathbb{P}[X \in A]$, which tells us how likely the value of X is to be within the set $A \subseteq \mathbb{R}^d$.

Definition 1.2.1 Let X and Y be random variables taking values in $\mathbb{R}^d$. We say that X and Y are *equal in distribution* if $\mathbb{P}[X \in A] = \mathbb{P}[Y \in A]$ for all $A \subseteq \mathbb{R}^d$. We write this relationship as $X \stackrel{\mathrm{d}}{=} Y$.

In the case $d = 1$ we also have the *cumulative distribution function* $F_X(x) = \mathbb{P}[X \leq x]$ which tells us how likely the value of X is to be less than or equal to $x \in \mathbb{R}$. For random variables X and Y taking values in $\mathbb{R}$,

$$F_X = F_Y \text{ if and only if } X \stackrel{\mathrm{d}}{=} Y. \tag{1.2}$$

We won't prove (1.2) within this course, although it is hopefully not surprising to you.

Example 1.2.2 It is important to understand that Definition 1.2.1 is not the same thing as equality. For example, let $X \sim \mathrm{N}(0,1)$ and let $Y = -X$. Then

$$\mathbb{P}[Y \leq x] = \mathbb{P}[-x \leq X] = \int_{-x}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = \mathbb{P}[X \leq x],$$

where we have made the substitution $z = -y$. Hence $F_X = F_Y$ so from (1.2) we have $X \stackrel{\mathrm{d}}{=} Y$. But $X = Y$ only happens when $X = Y = 0$, which has probability zero.

A perhaps simpler example: if X and Y are independent $\mathrm{N}(0,1)$ random variables then $X \stackrel{\mathrm{d}}{=} Y$, but $\mathbb{P}[X = Y] = \mathbb{P}[X - Y = 0]$ and $X - Y \sim \mathrm{N}(0, 1+1) \stackrel{\mathrm{d}}{=} \mathrm{N}(0,2)$ so $\mathbb{P}[X = Y] = 0$.

Note that we have used the notation $\sim \mathrm{N}(0,1)$ in Example 1.2.2. We might wonder what the difference between the symbols $\sim$ and $\stackrel{\mathrm{d}}{=}$ is. Formally, they have the same meaning, but we tend to use $\sim$ when we are referring to a named distribution, and $\stackrel{\mathrm{d}}{=}$ when we are comparing two existing random variables. That is a convention and not a rule, so you can use $\sim$ and $\stackrel{\mathrm{d}}{=}$ interchangeably if you wish.

©Nic Freeman, University of Sheffield, 2024.

1.2.1 Identifying distributions

In this section we give some results that help to identify the relationship $X \stackrel{\mathrm{d}}{=} Y$. Note that this also helps us identify when random variables have named distributions. The discrete case is easily dealt with.

Lemma 1.2.3 *Suppose that X and Y are discrete random variables. Then $X \stackrel{\mathrm{d}}{=} Y$ if and only if $p_X = p_Y$.*

Note that the statement $p_X = p_Y$ means that the functions p_X and p_Y are equal, that is $p_X(x) = p_Y(x)$ for all $x \in \mathbb{R}^d$. The proof is left for you in Problem 1.9.

The situation for continuous random variables is a bit more complicated. If X and Y are continuous random variables with $f_X = f_Y$, then it is clear from (1.1) that $X \stackrel{\mathrm{d}}{=} Y$, but it is possible to have $X \stackrel{\mathrm{d}}{=} Y$ and for f_X and f_Y to be 'different in an unimportant way'. You should have already seen examples of this situation, like the following.

Example 1.2.4 The probability density functions

$$f_X(x) = \begin{cases} 1 & \text{for } x \in (0,1) \\ 0 & \text{otherwise,} \end{cases} \qquad f_Y(y) = \begin{cases} 1 & \text{for } y \in [0,1] \\ 0 & \text{otherwise,} \end{cases}$$

define random variables X and Y. Note that $f_X(x) = f_Y(x)$ for all $x \in \mathbb{R}$ except for $x = 0$ and $x = 1$, so f_X and f_Y are different, but only very slightly! You might think of these as the continuous uniform distributions $X \sim \text{Uniform}((0,1))$ and $Y \sim \text{Uniform}([0,1])$, but they are really the same distribution because $\mathbb{P}[X = 0] = \mathbb{P}[X = 1] = \mathbb{P}[Y = 0] = \mathbb{P}[Y = 1] = 0$.

We need to handle this point carefully because, in this course, we don't assume enough mathematical background to explain precisely what we mean by 'different in an unimportant way'. We do need to know the following facts, however:

- For a random variable X with range in $\mathbb{R}$, changing the value of $f_X(x)$ on a finite set of $x \in \mathbb{R}$ will not change the distribution of X (as in Example 1.2.4).
- For a random variable X with range in $\mathbb{R}^2$, the same is true, but we can also change the value of $f_X(x)$ on a finite set of lines (in $\mathbb{R}^2$) without changing the distribution of X.

  Similar things work in higher dimensions too, but we won't need those.

We sometimes think of random variables as being defined by probability mass functions or probability density functions. This is a slight abuse of terminology: as we have discussed above, the p.m.f. and p.d.f. specify the distribution. If you are asked to 'find' the random variable X, or to find the distribution of X, then a statement of the p.m.f. or p.d.f. will suffice. You should always specify the range of values for which the p.m.f. of p.d.f. is non-zero.

©Nic Freeman, University of Sheffield, 2024.

1.2.2 Normalizing constants

Often you will find that the p.m.f. or p.d.f of some random variable X appears in the form

$$\mathbb{P}[X = x] = \frac{1}{Z}g(x) \qquad \text{or} \qquad f_X(x) = \frac{1}{Z}g(x) \tag{1.3}$$

where Z does not depend on x. In such cases Z is known as a *normalizing constant*. Its role is to make sure that p.m.f. sums (over $x \in R_X$) to one, and the p.d.f. integrates (again, over $x \in R_X$) to one. We have written $\frac{1}{Z}$ because normalizing constants often appear in a denominator e.g. $\frac{1}{\sqrt{2\pi}}$ in $f_{\mathrm{N}(0,1)}(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$, but they don't have to appear that way up e.g. λ in $f_{\mathrm{Exp}(\lambda)}(x) = \lambda e^{-\lambda x}$.

Lemma 1.2.5 *Suppose that X and Y are random variables.*

1. *If X and Y are discrete, with probability mass functions in the form $p_X(x) = \frac{1}{Z}g(x)$ and $p_Y(x) = \frac{1}{Z'}g(x)$ then $X \overset{\mathrm{d}}{=} Y$.*
2. *If X and Y are continuous, with probability density functions in the form $f_X(x) = \frac{1}{Z}g(x)$ and $f_Y(x) = \frac{1}{Z'}g(x)$ then $X \overset{\mathrm{d}}{=} Y$.*

Proof: (⊘) Note that in (1.3) the normalizing constant Z is determined by $g(x)$; in the discrete case we have $Z = \sum_{x \in R} g(x)$ and in the continuous case we have $Z = \int_{\mathbb{R}} g(x)\,dx$. Hence in both cases, the fact that X and Y are random variables implies that $Z = Z'$. The lemma now follows from Lemma 1.2.3 for the discrete case, and from our discussion below Lemma 1.2.3 for the continuous case. ■

Example 1.2.6 If $X \sim \mathrm{Gamma}(\alpha, \beta)$ then $f_X(x) = \frac{\beta^\alpha}{\Gamma(a)}x^{\alpha-1}e^{-\beta x}$ for $x > 0$, where $\Gamma : (0, \infty) \to (0, \infty)$ is the Γ-function. By Lemma 1.2.5, if Y is any other random variable in the form $f_Y(y) = (\text{constant}) \times x^{\alpha-1}e^{-\beta x}$, then we have $Y \sim \mathrm{Gamma}(\alpha, \beta)$.

When we have $p_X(x) = \frac{1}{Z}g(x)$ and $p_Y(x) = \frac{1}{Z'}g(x)$, as in part 1 of Lemma 1.2.5, it is common to summarize this relationship as $p_X \propto p_Y$. In words, p_X *is proportional to* p_Y. The same applies to part 2 of Lemma 1.2.5. This notation can save time, but we will avoid using it while we are focused on understanding conditioning and Bayesian models in Chapters 1-3. We will begin to use it in Section 4.1, where it will become very helpful in keeping our calculations tidy, and we will discuss it further at that point.

©Nic Freeman, University of Sheffield, 2024.

1.3 Families of random variables

Definition 1.3.1 We use the term *family (of distributions)* with *parameter space* $\Pi \subseteq \mathbb{R}^d$ to mean that each $\theta \in \Pi$ corresponds to a random variable M_θ, with parameters given by θ. We require that all the random variables within a given family have the same range R, which we call the range of the family.

For example:

- The *Beta family* refers to the distributions $\text{Beta}(\alpha, \beta)$ where the parameter $\theta = (\alpha, \beta)$ takes values in parameter space $\Pi = (0, \infty)^2$. It has range $[0, 1]$.
- The *Binomial family* refers to the distributions $\text{Bin}(n, p)$ where the parameter $\theta = (n, p)$ takes values in parameter space $\Pi = \mathbb{N} \times [0, 1]$. It has range $\mathbb{N}$.

We say that a family is *discrete* if it is made up of (exclusively) discrete random variables, and *continuous* if is it made up of (exclusively) continuous random variables. So the Beta family is an continuous family and the Binomial family is a discrete family.

We will use this term for statistical models, written $(M_\theta)_{\theta \in \Pi}$ with parameter θ. For this reason we will often refer to (M_θ) as a *model family*.

Assumption 1.3.2 For our model families, we require that $\theta \mapsto \mathbb{P}[M_\theta \in A]$ is a continuous function, for all $A \subseteq R$.

The purpose of this assumption is that if we change θ slightly then we only change the distribution of M_θ slightly. This will be necessary for our inference methods later on. This condition holds for most common families of random variables, including all of those listed on the reference sheets in Appendix A.

©Nic Freeman, University of Sheffield, 2024.

1.4 Conditioning on location

Lemma 1.4.1 *Let X be a random variable, and let $A \subseteq \mathbb{R}$ be such that $\mathbb{P}[X \in A] > 0$. Then there exists a random variable Y such that*

1. $\mathbb{P}[Y \in A] = 1$.
2. $\mathbb{P}[Y \in B] = \frac{\mathbb{P}[X \in B]}{\mathbb{P}[X \in A]}$ *for all* $B \subseteq A$. (1.4)

If Y' is any random variable satisfying these two conditions then $Y \overset{\mathrm{d}}{=} Y'$.

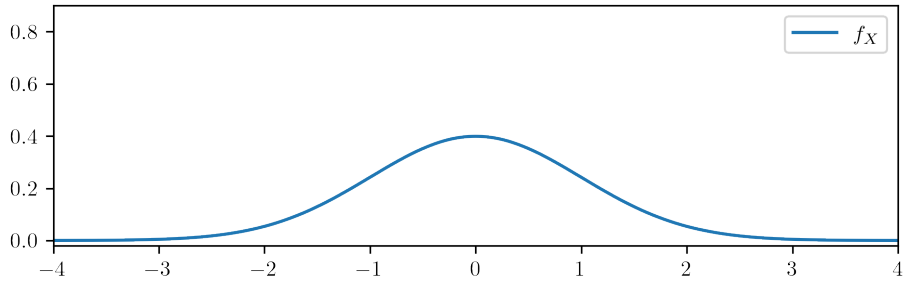
Definition 1.4.2 The distribution of Y is known as the '*conditional distribution of X given the event $\{X \in A\}$*'. We write this relationship in symbols as $Y \overset{\mathrm{d}}{=} X|_{\{X \in A\}}$.

Note that we use $\overset{\mathrm{d}}{=}$ and not $=$ in Definition 1.4.2. It is possible to make lots of different random variables Y that satisfy properties 1 and 2 of Lemma 1.4.1, but they all have the same distribution. Using $\overset{\mathrm{d}}{=}$ instead of $=$ captures this fact. We will prove Lemma 1.4.2 shortly, but let us first concentrate on getting the intuition right. Property 1 in Lemma 1.4.1 says that Y is always inside the set A. Property 2 says that, inside A, Y behaves like X. (Taking $B = A$ in property 2 gives property 1, but it will be helpful to refer back to them separately.)

You might like to think of Y as what happens if the random variable X is forced to sit inside the set A. It is still (in general) a random quantity, and it reflects exactly the random behaviour of X inside the set A, but all the behaviour of X outside of A is forgotten. Another way to understand Y is via *rejection sampling*: we could repeatedly take samples of X until we obtain a sample of X that is inside the set A. The random quantity that we obtain from this procedure has precisely the same behaviour as Y.

We will use the usual language of probability to rewrite the event $\{X \in A\}$ when it is more intuitive to do so. For example, if $A = [a, b]$ then we might write $X|_{\{X \in [a,b]\}}$, or if $A = \{a\}$ then we might write $X|_{\{X=a\}}$.

Example 1.4.3 Suppose that $X \sim \mathrm{N}(0, 1)$. The values taken by X are spread out across the real line. The probability density function of X, given by $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ allows us to visualize the random location of X:



The random variable X is more likely to be in locations where $f_X(x)$ takes larger values, or more precisely $\mathbb{P}[a \leq X \leq b] = \int_a^b f_X(x)\, dx$, the area under the curve f_X between a and b.

Let $Y = X|_{\{X \in [0,\infty)\}}$. We can use the properties given in Lemma 1.4.1 to find the distribution of Y, where we take $A = [0, \infty)$. Firstly, note that property 1 gives $\mathbb{P}[Y \leq 0] = 0$, so $\mathbb{P}[Y \leq y] = 0$ for all $y \leq 0$. For $y > 0$ we have

$$\mathbb{P}[Y \leq y] = \mathbb{P}[Y \leq 0] + \mathbb{P}[0 \leq Y \leq y]$$

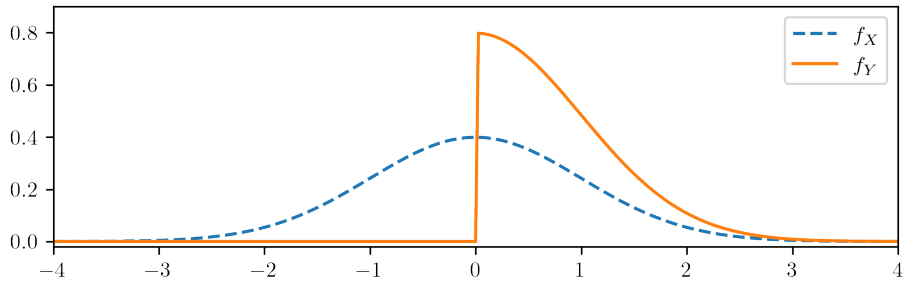
©Nic Freeman, University of Sheffield, 2024.

$$\begin{aligned} &= 0 + \frac{\mathbb{P}[0 \leq X \leq y]}{\mathbb{P}[0 \leq X < \infty]} \\ &= \frac{\int_0^y f_X(x)\,dx}{1/2} \\ &= \int_0^y 2\frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx. \end{aligned} \tag{1.5}$$

We therefore obtain that Y is a continuous random variable with p.d.f.

$$f_Y(y) = \begin{cases} 0 & \text{for } y \leq 0 \\ \sqrt{\frac{2}{\pi}} e^{-y^2/2} & \text{for } y > 0. \end{cases}$$

Plotting f_X and f_Y on the same axis we obtain



Note that $f_Y(y)$ is zero for $y \in (-\infty, 0)$, and double the value of $f_X(y)$ on $y \in [0, \infty)$. This is not a coincidence. It occurs precisely because Y retains the 'same' randomness as X inside the set $A = [0, \infty)$, but Y can only take values inside the set A, so the p.d.f. within in that segment must be scaled up to ensure that $\int_{\mathbb{R}} f_Y(y)\,dy = 1$. That fact that this scaling up is by a factor of 2, in this example, comes from (1.5) and in particular from $\mathbb{P}[X \in A] = \frac{1}{2}$.

In Problem **1.7** you can explore Example 1.4.3 a bit further, and there is a more general version in Problem **1.8**. We will now give a proof of Lemma 1.4.1, based on the idea of rejection sampling that we mentioned below Lemma 1.4.1.

Proof of Lemma 1.4.1: (⊘) We will discuss this proof in lectures, because Lemma 1.4.1 is fundamental to the whole course and the key ideas of the proof are helpful to understand. The details are less important to us: our focus is on becoming competent practitioners of Bayesian statistics, rather than on becoming able to develop its theory.

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. copies of the random variable X. Let $N \in \mathbb{N}$ be the number of the first copy for which $X_N \in A$. Let us write $q = \mathbb{P}[X_n \in A] > 0$. The events $\{X_n \in A\}$ are a sequence of independent trials with success probability q, so the number of trials until the first success is Geometric(q). In particular, since $\mathbb{P}[\text{Geometric}(q) < \infty] = 1$, a success will eventually happen. Let $N = \min\{n \in \mathbb{N};\, X_n \in A\}$ be the number of trials until this first success.

We claim that $Y = X_N$ satisfies the two properties required in the statement of the lemma. By definition of N we have $\mathbb{P}[X_N \in A] = 1$, which shows the first property. To see the second, for $B \subseteq A$ we have

$$\mathbb{P}[X_N \in B] = \sum_{n=1}^{\infty} \mathbb{P}[X_N \in B \text{ and } N = n]$$

©Nic Freeman, University of Sheffield, 2024.

$$
\begin{aligned}
&= \sum_{n=1}^{\infty} \mathbb{P}[X_n \in B \text{ and } X_{n-1} \notin A, X_{n-2} \notin A, \ldots, X_1 \notin A] \\
&= \sum_{n=1}^{\infty} \mathbb{P}[X_n \in B] \times \mathbb{P}[X_{n-1} \notin A] \times \ldots \times \mathbb{P}[X_1 \notin A] \\
&= \sum_{n=1}^{\infty} \mathbb{P}[X_n \in B](1-q)^{n-1} \\
&= \mathbb{P}[X \in B] \frac{1}{1-q} \sum_{n=1}^{\infty} (1-q)^n \\
&= \mathbb{P}[X \in B] \frac{1}{1-q} \frac{1-q}{q} \\
&= \frac{\mathbb{P}[X \in B]}{\mathbb{P}[X \in A]}.
\end{aligned}
$$

In the above we use independence of the X_n to deduce the third line. To deduce the fifth line we use that X_n and X have the same distribution, hence $\mathbb{P}[X_n \in B] = \mathbb{P}[X \in B]$, and to deduce the final line we use this same fact with the definition of q. Hence, $Y = X_N$ satisfies the properties required.

To prove the final part of the lemma, note that if Y and Y' both satisfy the two properties then for any $C \subseteq \mathbb{R}$ we have

$$
\mathbb{P}[Y \in C] = \mathbb{P}[Y \in C \cap (\mathbb{R} \setminus A)] + \mathbb{P}[Y \in C \cap A] = 0 + \frac{\mathbb{P}[X \in C \cap A]}{\mathbb{P}[X \in A]} \tag{1.6}
$$

and the same holds for Y'. To deduce the last equality in (1.6) we have used property 1 for the first term and property 2 for the second term. The right hand side of (1.6) only depends on X, hence $\mathbb{P}[Y \in C] = \mathbb{P}[Y' \in C]$. Thus $Y \overset{\mathrm{d}}{=} Y'$. ■

Lemma 1.4.4 *Let X be a random variable taking values in $\mathbb{R}^d$.*

1. *Let $a \in \mathbb{R}^d$ be such that $\mathbb{P}[X = a] > 0$. Then $X|_{\{X=a\}} \overset{\mathrm{d}}{=} a$.*
2. *It holds that $X|_{\mathbb{R}^d} \overset{\mathrm{d}}{=} X$.*

Proof: We should think of these two cases as (1) conditioning on 'X is equal to a' and (2) conditioning on 'X could be anywhere'. The results of doing so should not be a surprise in either case!

To prove the part 1, if we put $A = \{a\}$ in Lemma 1.4.1 then the first property gives $\mathbb{P}[X|_{\{X=a\}} = a] = 1$. Therefore $X|_{\{X=a\}}$ and a are equal (with probability 1) which means they have the same distribution.

To prove part 2, put $A = \mathbb{R}^d$ in Lemma 1.4.1, then the second property gives $\mathbb{P}[Y \in B] = \frac{\mathbb{P}[X \in B]}{\mathbb{P}[X \in \mathbb{R}^d]} = \frac{\mathbb{P}[X \in B]}{1} = \mathbb{P}[X \in B]$ for all $B \subseteq \mathbb{R}^d$, which tells us that $\mathbb{P}[X|_{\mathbb{R}^d} \in B] = \mathbb{P}[X \in B]$. That is, $X|_{\mathbb{R}^d}$ and X have the same distribution. ■

©Nic Freeman, University of Sheffield, 2024.

1.5 Conditioning and correlations

This section demonstrates the effect of taking a jointly distributed random variable $(X, Y) \in \mathbb{R}^2$, and conditioning X to be within a particular location. If X and Y are independent then conditioning X will have no effect on Y (see Exercise 1.10 for details), but if they are dependent then conditioning on the location of X will *change* the distribution of the y coordinate. This is because X and Y affect each other so, if we force X to do something, it will also have some effect on Y.

Let us introduce some notation for these ideas.

- We write $(X, Y)|_{\{X \in A\}}$ as a shorthand for $(X, Y)|_{\{(X,Y) \in A \times \mathbb{R}^d\}}$, where we restrict the location of X (to be inside A) but we do not restrict the location of Y (because $Y \in \mathbb{R}^d$ is true anyway).
- We write $Y|_{\{X \in A\}}$ for the y coordinate of $(X, Y)|_{\{X \in A\}}$.

In this notation, we idea we described above is that, if X and Y are dependent, the random variables Y and $Y|_{\{X \in A\}}$ will have different distributions.

Lemma 1.5.1 *Let X and Y be random variables, with $A \subseteq R_X$, $B \subseteq R_Y$ and $\mathbb{P}[X \in A] > 0$. Then*

$$\mathbb{P}[Y|_{\{X \in A\}} \in B] = \frac{\mathbb{P}[X \in A, Y \in B]}{\mathbb{P}[X \in A]}. \tag{1.7}$$

Proof: From part 2 of Lemma 1.4.1,

$$\mathbb{P}[(X, Y)|_{\{X \in A\}} \in A \times B] = \frac{\mathbb{P}[(X, Y) \in A \times B]}{\mathbb{P}[(X, Y) \in A \times \mathbb{R}^d]} = \frac{\mathbb{P}[X \in A, Y \in B]}{\mathbb{P}[X \in A]}. \tag{1.8}$$

Part 1 of Lemma 1.4.1 tells us that $(X, Y)|_{\{X \in A\}}$ has range $A \times R_Y$. Hence $X|_{\{X \in A\}}$ has range A, which means that $\mathbb{P}[Y|_{\{X \in A\}} \in B] = \mathbb{P}[(X, Y)|_{\{X \in A\}} \in A \times B]$. Combining this fact with (1.8) completes the proof. ■

If X and Y are discrete then taking $A = \{x\}$ and $B = \{y\}$ gives us $\mathbb{P}[Y|_{\{X=x\}} = y] = \frac{\mathbb{P}[X=x, Y=y]}{\mathbb{P}[X=x]}$. This formula, and more generally (1.7), should feel familiar. In earlier courses you will probably have seen equations like $\mathbb{P}[Y = y \,|\, X = x] = \frac{\mathbb{P}[X=x, Y=y]}{\mathbb{P}[X=x]}$, which has essentially the same meaning but different notation. The reason for introducing $Y|_{\{X=A\}}$ is simply that it is easier to understand probability when we can imagine random objects.

Example 1.5.2 Suppose that we roll a fair dice, with outcomes $Z = 1, 2, \ldots, 6$. Define the random variables

$$X = \begin{cases} 1 & \text{if } Z \text{ is odd,} \\ 0 & \text{if } Z \text{ is even,} \end{cases} \qquad Y = \begin{cases} 0 & \text{if } Z \leq 3, \\ 1 & \text{if } Z \geq 4, \end{cases}$$

which are dependent. We can illustrate their joint distribution with a table of values:

Z	1	2	3	4	5	6
X	1	0	1	0	1	0
Y	0	0	0	1	1	1
$\mathbb{P}$[column]	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

©Nic Freeman, University of Sheffield, 2024.

Each column is a possible outcome, each of which has probability $\frac{1}{6}$. Conditioning on the event $X = 1$ forces the outcome to be within the shaded columns.

The distribution of Y is easily found:

$$\mathbb{P}[Y = 0] = \mathbb{P}[Z \in \{1, 2, 3\}] = \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{2},$$
$$\mathbb{P}[Y = 1] = \mathbb{P}[Z \in \{4, 5, 6\}] = \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{2}.$$

By Lemma 1.5.1 we have

$$\mathbb{P}[Y|_{\{X=1\}} = 0] = \frac{\mathbb{P}[Y = 0, X = 1]}{\mathbb{P}[X = 1]} = \frac{\mathbb{P}[Z \in \{1, 3\}]}{\mathbb{P}[Z \in \{1, 3, 5\}]} = \frac{\frac{1}{6} + \frac{1}{6}}{\frac{1}{6} + \frac{1}{6} + \frac{1}{6}} = \frac{2}{3}$$

and so $\mathbb{P}[Y|_{\{X=1\}} = 1] = 1 - \frac{2}{3} = \frac{1}{3}$. As we expected, the distributions of Y and $Y|_{\{X=1\}}$ are different.

Compare the situation of Example 1.5.2 to that of a random variable (X, Y) taking values in $\mathbb{R}^n \times \mathbb{R}^d \equiv \mathbb{R}^{n+d}$. Here, X takes values in $\mathbb{R}^n$ and Y takes values in $\mathbb{R}^d$. If X and Y are dependent, then we should still expect that conditioning one affects the distribution of the other. This is the key fact that we take away from this section.

Remark 1.5.3 In a multivariate situation, say (X, Y_1, Y_2), we could do something similar and find the conditional distributions of $(Y_1, Y_2)|_{\{X \in A\}}$ as well as $Y_1|_{\{X \in A\}}$ and $Y_2|_{\{X \in A\}}$. A slightly subtle point is that

$$(Y_1, Y_2)|_{\{X \in A\}} \overset{\mathrm{d}}{=} \left(Y_1|_{\{X \in A\}}, Y_2|_{\{X \in A\}}\right).$$

In words, conditioning two coordinates on the same event, is equivalent to conditioning each coordinate individually on that event – as we would intuitively expect.

We will use this fact later on when we work with multivariate situations, but we won't include a proof within our course. It isn't difficult to check, but it wouldn't help us understand anything more than we already do.

©Nic Freeman, University of Sheffield, 2024.

1.6 Conditioning on events with zero probability

Lemma 1.4.1 requires that the event A, on which we condition, has positive probability. Without that condition equation (1.4) would become $\frac{0}{0}$, which is undefined. Despite this problem it is often possible to make sense of the result of conditioning on an event of probability zero. The mathematical theory here is much more complicated than we can cover, so instead we will explain the key idea and focus on a small set of well-behaved situations.

We will take a random variable (Y, Z) and condition Z on the event $\{Y = y\}$, where Y is continuous. In this case $\mathbb{P}[Y = y] = 0$. The key idea is this: if there exists a random variable Z^* such that

$$\mathbb{P}\left[Z|_{\{|Y-y|\leq\epsilon\}} \in A\right] \to \mathbb{P}[Z^* \in A] \qquad \text{as } \epsilon \to 0 \tag{1.9}$$

for all $A \subseteq \mathbb{R}^n$, then we extend Definition 1.4.2; we say that Z^* is the conditional distribution of Z given $\{Y = y\}$, written $Z^* \stackrel{\mathrm{d}}{=} Z|_{\{Y=y\}}$. There are many examples of random variables (Y, Z) for which the limit in (1.9) does not exist, or fails to behave like a conditional probability. There are, also, many examples where (1.9) results in a random variable Y that behaves exactly like we would expect, based on the intuition we have built up from Lemma 1.4.1.

Here is a case where it does work, that you've seen before. In earlier courses you may have been told that (1.10) was the definition of a conditional p.d.f., but that is not entirely honest. It is a consequence of (1.9) and it requires some conditions.

Lemma 1.6.1 *Let (Y, Z) be continuous random variables, where Y takes values in $\mathbb{R}^n$ and Z takes values in $\mathbb{R}^d$. Suppose that $f_Y(y) > 0$ and that both $f_Y(y)$ and $f_{Y,Z}(y, z)$ are continuous functions. Then $Z|_{\{Y=y\}}$ is a continuous random variable with p.d.f.*

$$f_{Z|_{\{Y=y\}}}(z) = \frac{f_{Y,Z}(y, z)}{f_Y(y)}. \tag{1.10}$$

Proof: (⊘) We will sketch an argument to show why the result holds, but we won't include a proof here. For simplicity, let us assume that Y and Z both takes values in $\mathbb{R}$. From Lemma 1.5.1 we have

$$\mathbb{P}[Z|_{\{|Y-y|\leq\epsilon\}} \in B] = \frac{\mathbb{P}[|Y - y| \leq \epsilon, Z \in B]}{\mathbb{P}[|Y - y| \leq \epsilon]} = \frac{\int_B \int_{y-\epsilon}^{y+\epsilon} f_{Y,Z}(u, z)\, du\, dz}{\int_{y-\epsilon}^{y+\epsilon} f_Y(u)\, du}.$$

Our continuity assumptions mean that when ϵ is small and $|u - y| \leq \epsilon$, we can approximate $f_{Y,Z}(u, z) \approx f_{Y,Z}(y, z)$ and $f_Y(u) \approx f_Y(y)$. Putting both these approximations in,

$$\mathbb{P}[Z|_{\{|Y-y|\leq\epsilon\}} \in B] \approx \frac{\int_B \int_{y-\epsilon}^{y+\epsilon} f_{Y,Z}(y, z)\, du\, dz}{\int_{y-\epsilon}^{y+\epsilon} f_Y(u)\, du} = \frac{2\epsilon \int_B f_{Y,Z}(y, z)\, dz}{2\epsilon f_Y(y)} = \int_B \frac{f_{Y,Z}(y, z)}{f_Y(y)}\, dz. \tag{1.11}$$

Because of our approximation the right hand side of (1.11) does not contain ϵ. This suggests that letting $\epsilon \to 0$ should result in $\lim_{\epsilon\to 0} \mathbb{P}[Z|_{\{|Y-y|\leq\epsilon\}} \in B] = \int_B \frac{f_{Y,Z}(y,z)}{f_Y(y)}\, dz$. Combining this formula with Definition 1.1.1 and (1.9) gives that $Z|_{\{Y=y\}}$ exists and is a continuous random variable, with the p.d.f. as claimed. ■

Example 1.6.2 Let (Y, Z) be a continuous random variable taking values in $\mathbb{R}^2$, with (joint) probability density function

$$f_{(Y,Z)}(y, z) = \frac{1}{\sqrt{2\pi^3}} \frac{e^{-y^2/2}}{1 + (z - y)^2}. \tag{1.12}$$

©Nic Freeman, University of Sheffield, 2024.

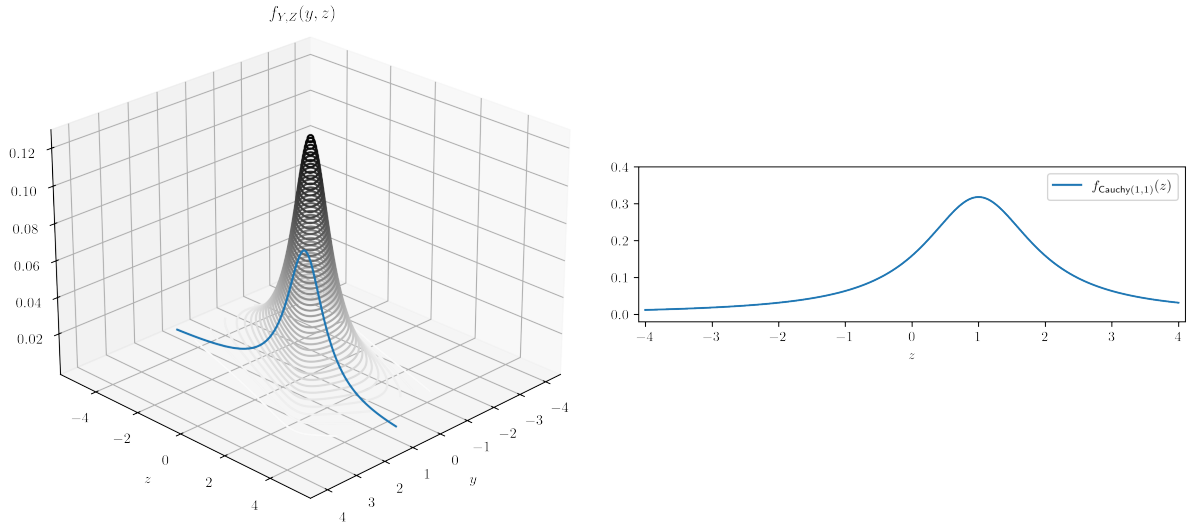
for all $y, z \in \mathbb{R}$. Note that this is a continuous function (or see the plot below). We can compute

$$\begin{aligned} f_Y(y) = \int_{\mathbb{R}} f_{Y,Z}(y,z)\,dz &= \frac{1}{\sqrt{2\pi^3}} e^{-y^2/2} \int_{-\infty}^{\infty} \frac{1}{1+(z-y)^2}\,dz \\ &= \frac{1}{\sqrt{2\pi^3}} e^{-y^2/2} \int_{-\infty}^{\infty} \frac{1}{1+z^2}\,dz \\ &= \frac{1}{\sqrt{2\pi^3}} e^{-y^2/2} [\arctan(z)]_{-\infty}^{\infty} \\ &= \frac{1}{\sqrt{2\pi^3}} e^{-y^2/2} \left(\frac{\pi}{2} - \frac{-\pi}{2} \right) \\ &= \frac{1}{\sqrt{2\pi}} e^{-y^2/2}, \end{aligned}$$

which we recognize as $Y \sim \mathrm{N}(0,1)$. We will condition on $\{Y = 1\}$. Clearly $f_Y(1) > 0$, so Lemma 1.6.1 applies, and we obtain

$$f_{Z|_{\{Y=1\}}}(y) = \frac{f_{Y,Z}(1,z)}{f_Y(1)} = \frac{\frac{1}{\sqrt{2\pi^3}} \frac{e^{-1}}{1+(z-1)^2}}{\frac{1}{\sqrt{2\pi}} e^{-1}} = \frac{1}{\pi} \frac{1}{1+(z-1)^2}$$

which we recognize as $Z|_{\{Y=1\}} \sim \text{Cauchy}(1,1)$. We can plot $f_{Y,Z}(y,z)$ and $f_{Z|_{\{Y=1\}}}$ as follows:



The line on the left hand picture corresponds to $y = 1$. You can see that it has the shape of the Cauchy$(1,1)$ p.d.f. in the right hand picture.

Remark 1.6.3 (⊘) It is possible to extend Lemma 1.6.1 to weaken the assumption of continuity. This requires some care and we won't explore the details here, although we will sometimes use (1.10) in these cases. As a general rule it is dangerous to condition when $f_{Y,Z}(y,z)$ features discontinuities that might influence the conditioning.

Remark 1.6.4 Taking $Y = Z$ in (1.9), a similar approximation argument to the proof of Lemma 1.6.1 shows that for a continuous random variable Y and $y \in \mathbb{R}_Y$, we have $Y|_{\{Y=y\}} \overset{\mathrm{d}}{=} y$. We already knew this for discrete random variables, in Lemma 1.4.4, so it is hopefully easy to believe. We record this fact because we will need it in Chapter 8.

©Nic Freeman, University of Sheffield, 2024.

1.7 Families with random parameters

In this section we are interested to take a model family $(M_\theta)_{\theta\in\Pi}$ and treat the parameter θ as a random variable, which will be denoted by a capital letter Θ. We think of first sampling the value of Θ and then (using whatever value we obtain) taking a sample X from M_Θ. We will have a detailed discussion of how this idea becomes useful in Section 2.1. For now let us note that it increases the range of models that we have available.

To make sense of the idea, let us state it more precisely. We want random variables X and Θ such that $X|_{\{\Theta=\theta\}} \overset{\mathrm{d}}{=} M_\theta$. In this section we show that a pair (X,Θ) with this property is given by the distribution

$$\mathbb{P}[X\in A,\Theta\in B] = \int_B \mathbb{P}[M_\theta\in A] f_\Theta(\theta)\,d\theta. \tag{1.13}$$

where (M_θ) is a family of distributions with range $R\subseteq\mathbb{R}^n$, as defined in Section 1.3, and f_Θ is a probability density function with range $\Pi\subseteq\mathbb{R}^d$. This is a type of random variable you may not have seen before. We will shortly show that the Θ part is a continuous random variable, but the X part might be discrete or continuous, depending on (M_θ).

Our notation strongly suggests that we expect f_Θ to be the (marginal) probability density function of Θ, and we can confirm this by setting $A=\mathbb{R}^n$, in which case equation (1.13) becomes $\mathbb{P}[\Theta\in B]=\int_B f_\Theta(\theta)$. We can also find the marginal distribution of X, by setting $B=\mathbb{R}^d$, giving

$$\mathbb{P}[X\in A] = \mathbb{P}[X\in A,\Theta\in\mathbb{R}^d] = \int_{\mathbb{R}^d} \mathbb{P}[M_\theta\in A] f_\Theta(\theta)\,d\theta,$$

but that formula doesn't really explain what is going on here. The relationship that we are interested in comes from the following lemma.

Lemma 1.7.1 *Let (M_θ) and (X,Θ) have distribution given by* (1.13)*. Suppose that $f_\Theta(\theta)>0$ and that $t\mapsto f_\Theta(t)$ is continuous at $t=\theta$. Then $X|_{\{\Theta=\theta\}}\overset{\mathrm{d}}{=}M_\theta$.*

Proof: (⊘) We give a sketch proof to illustrate the idea, in similar style to Lemma 1.6.1. From Lemma 1.5.1, for $A\in\mathbb{R}^d$ we have

$$\mathbb{P}[X|_{\{|\Theta-\theta|\le\epsilon\}}\in A] = \frac{\mathbb{P}[|\Theta-\theta|\le\epsilon, X\in A]}{\mathbb{P}[|\Theta-\theta|\le\epsilon]} = \frac{\int_{\theta-\epsilon}^{\theta+\epsilon}\mathbb{P}[M_t\in A]f_\Theta(t)\,dt}{\int_{\theta-\epsilon}^{\theta+\epsilon} f_\Theta(t)\,dt}$$

The second equality follows from (1.13) with $B=[\theta-\epsilon,\theta+\epsilon]$, for the numerator, and from the fact that f_Θ is the p.d.f. of Θ, for the denominator. Using continuity, from the statement of the lemma and from Assumption 1.3.2, for $|\theta-t|\le\epsilon$ we can approximate $f_\Theta(t)\approx f_\Theta(\theta)$ and $\mathbb{P}[M_t\in A]\approx\mathbb{P}[M_\theta\in A]$. This gives

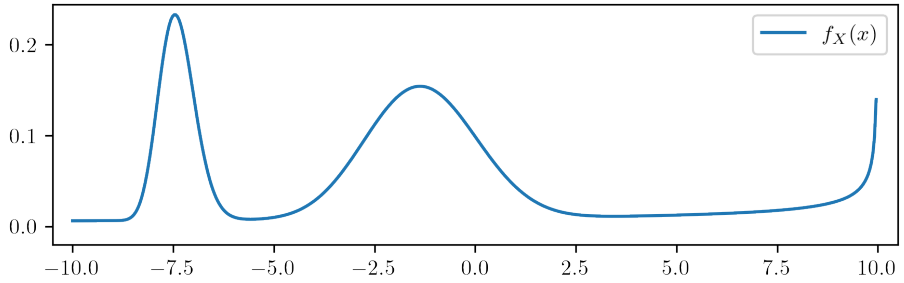
$$\mathbb{P}[X|_{\{|\Theta-\theta|\le\epsilon}\in A] \approx \frac{\int_{\theta-\epsilon}^{\theta+\epsilon}\mathbb{P}[M_\theta\in A]f_\Theta(\theta)\,dt}{\int_{\theta-\epsilon}^{\theta+\epsilon} f_\Theta(\theta)\,dt} = \frac{2\epsilon f_\Theta(\theta)\mathbb{P}[M_\theta\in A]}{2\epsilon f_\Theta(\theta)} = \mathbb{P}[M_\theta\in A].$$

Letting $\epsilon\to 0$ we have $\mathbb{P}[X|_{\{|\Theta-\theta|\le\epsilon}\in A]\to\mathbb{P}[M_\theta\in A]$, so by Definition 1.2.1 and (1.9) we have that $X|_{\{\Theta=\theta\}}$ is well defined and $X|_{\{\Theta=\theta\}}\overset{\mathrm{d}}{=}M_\theta$. ■

©Nic Freeman, University of Sheffield, 2024.

1.8 Exercises on Chapter 1

⋆ **1.1** (a) Consider a continuous random variable X with range $R_X = (-10, 10)$ and p.d.f. f_X sketched as follows:



In words, explain which parts of R_X are the most likely locations for a random sample of X to be.

(b) Let $\theta > 1$. Check that the function $f : \mathbb{R} \to [0, \infty)$ by $f(x) = \begin{cases} (\theta-1)x^{-\theta} & \text{for } x \geq 1 \\ 0 & \text{otherwise} \end{cases}$ is a probability density function i.e. check that integrating over its range gives 1.

⋆ **1.2** Let $X \sim \mathrm{N}(0, 1)$ and let Y be given by $\mathbb{P}[Y = 0] = \mathbb{P}[Y = 1] = \frac{1}{2}$, independently of Z. Define a random variable

$$Z = \begin{cases} 0 & \text{if } Y = 0, \\ X & \text{if } Y = 1. \end{cases}$$

Is Z a discrete random variable, a continuous random variable, or neither?

⋆⋆ **1.3** Let $a < b < c$, all elements of $\mathbb{R}$. Suppose that U has the continuous uniform distribution on $[a, c]$ and let $U' \overset{\mathrm{d}}{=} U|_{\{U \in [a,b]\}}$.

(a) Write out the meaning of $U' \overset{\mathrm{d}}{=} U|_{\{U \in [a,b]\}}$, in words.

(b) Show that U' has the continuous uniform distribution on $[a, b]$.

(c) Plot the probability density functions of U and U' on the same axis, for the case $a = 1, b = 2, c = 4$.

⋆⋆ **1.4** Suppose that G has the Geometric(p) distribution, that is $\mathbb{P}[G = g] = p^g(1 - p)$ for $g \in \{0, 1, \ldots\}$, where $p \in [0, 1]$. Let $G' \overset{\mathrm{d}}{=} G|_{\{G \geq n\}}$, where $n \in \mathbb{N}$.

(a) Find the distribution of G'.

(b) Consider the following claim.

> *Recall that G represents the time of the first failure in a sequence of independent trials, each of which has failure probability p. Conditioning G to be greater than or equal to n has the effect of forcing the first n trials to be successful, without changing the distribution of the remaining trials.*

Do you agree with this statement? Give brief reasons for your answer.

©Nic Freeman, University of Sheffield, 2024.

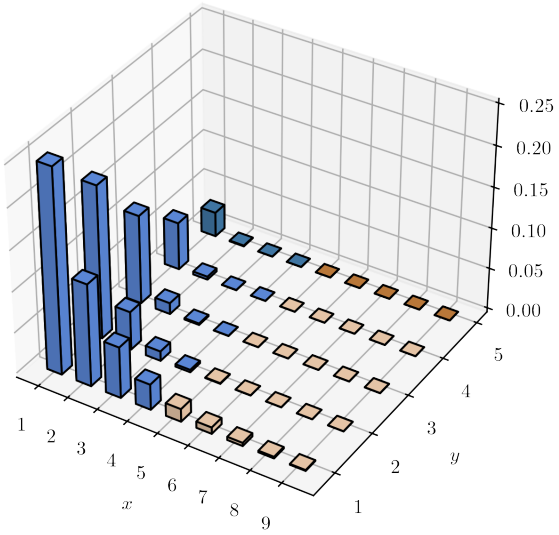
⋆ **1.5** Let $X \sim \text{Cauchy}(0,1)$ and let $x \in \mathbb{R}$. In the files `1_rejection_sampling.ipynb` and `1_rejection_sampling.Rmd` you can find a rejection sampling algorithm (as described in our comments below Lemma 1.4.1 or within the proof of that lemma) that obtains samples from $X|_{\{X \geq k\}}$.

Record the time it takes for the code to obtain some samples, for $k = 2^n$ for various values of $n \in \{1, 2, \ldots\}$. What do you notice? Repeat the experiment with some other distributions.

⋆⋆ **1.6** Let (X, Y) be a discrete random variable with values in $\mathbb{R}^2$ (so $n = d = 1$) with distribution given by

$$\mathbb{P}[(X,Y) = (x,y)] = \begin{cases} 2^{-xy}(1 - 2^{-y}) & \text{for } x, y \in \mathbb{N} \\ 0 & \text{otherwise.} \end{cases} \tag{1.14}$$

Here is a plot of this function, with shading to indicate regions for which $x \geq 5$ and/or $y = 5$.



(a) Check that the function given in (1.14) is a probability mass function with range $\mathbb{N}^2$ i.e. check that summing over its range gives 1. Are X and Y independent?

(b) (i) Show that for $y \in \mathbb{N}$ we have $\mathbb{P}[Y = y] = \left(\frac{1}{2}\right)^y$.

(ii) Show that for $x \in \mathbb{N}$ we have $\mathbb{P}[X|_{\{Y=5\}} = x] = \left(1 - \frac{1}{2^5}\right)\left(\frac{1}{2^5}\right)^{x-1}$.

(iii) Show that for $y \in \mathbb{N}$ we have $\mathbb{P}[Y|_{\{X \geq 5\}} = y] = \left(1 - \frac{1}{2^5}\right)\left(\frac{1}{2^5}\right)^{y-1}$.

⋆⋆ **1.7** Let $X \sim \text{N}(0,1)$ and set $A = [0, \infty)$, as in Example 1.4.3. Let $Y' = |X|$. Show that $Y' \overset{\text{d}}{=} X|_{\{X \in A\}}$.

⋆⋆⋆ **1.8** Let X be a continuous random variable and let $A \subseteq R_X$ with $\mathbb{P}[X \in A] > 0$. Show that $X|_{\{X \in A\}}$ is a continuous random variable with p.d.f.

$$f_{X|_{\{X \in A\}}}(x) = \begin{cases} \frac{f_X(x)}{\mathbb{P}[X \in A]} & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases}$$

⋆⋆⋆ **1.9** Prove Lemma 1.2.3.

⋆⋆⋆ **1.10** Suppose that X and Y are independent random variables, and that $\mathbb{P}[X \in A] > 0$. Show that $Y \overset{\text{d}}{=} Y|_{\{X \in A\}}$.

©Nic Freeman, University of Sheffield, 2024.

Chapter 2

Discrete Bayesian models

A *model* is a machine for simulating data. We hope that the data we simulate is a realistic, or approximately realistic, copy of some part of the real world. In probability and statistics we use *random* models, meaning that if we use the same model twice we won't generate the same data.

The reason for using a random model is that we can express our level of certainty. We don't know exactly what will happen, so instead we create a model that includes a range of possibilities, along with a description how likely these different possibilities are. If we have a high level of certainty then we might choose a model in which the likely outcomes are very similar to each other. If not, we might choose a model in which the likely outcomes span a wide range of possibilities.

We usually allow our models to have parameters. Parameters are 'input' values that we can use to control how a model behaves, and we try to choose them in a way that makes our model best reflect reality.

©Nic Freeman, University of Sheffield, 2024.

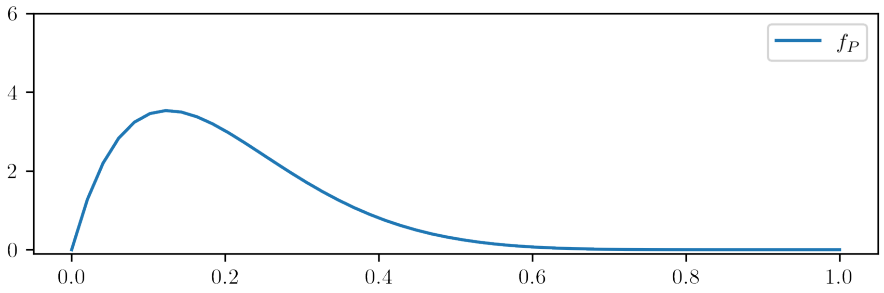
2.1 Models with random parameters

We are often confident in our choice of model, but we are almost never certain what the best choice of parameters is. In that situation, we usually do have some degree of confidence in which parameter values should be used. How should we communicate that information? One way is to use *random variables* for the parameters, as we did in Section 1.7. If we have a high level of certainty in which parameter values should be used, then we will choose a random variable in which the likely outcomes are all very similar parameter values. If not then we choose a random variable in which the likely outcomes span a wide range of different parameters. The 'parameters' become just another part of the random model.

Example 2.1.1 We want to model the random number X of successful experiments, out of a sequence of 10 independent experiments, where each experiment has the same probability $p \in [0,1]$ of success. A natural form for X is $\mathrm{Bin}(10,p)$, but we don't know the value of p, so let us set $M_p \sim \mathrm{Bin}(10,p)$. Then (M_p) is a family of distributions. We hope that one of these distributions will be a good model.

We are told (by a scientist colleague, say) that a reasonable estimate for p is $p \approx \frac{1}{5}$, but they don't seem very confident and for now that is all we know. We decide that our model for p will be a random variable $P \sim \mathrm{Beta}(2,8)$, which has $\mathbb{E}[P] = \frac{1}{5}$ and probability density function

$$f_P(p) = \begin{cases} 72p(1-p)^7\,dp & \text{for } p \in [0,1] \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$



The decision here to use $\mathrm{Beta}(2,8)$ is a bit arbitrary – we will spend a large part of this course thinking about how to make such choices well! It is only meant here as example of the type of model that we will be interested in. We have chosen a distribution with range $[0,1]$, which is the set of possible values for p, and with most of its mass within the region $p \approx \frac{1}{5}$ that was suggested to us.

The random variable (X,P) that results from this procedure is precisely the type we studied in Section 1.7. We want $X|_{\{P=p\}} \sim \mathrm{Bin}(10,p)$ and (for now) our best guess for the parameter is $P \sim \mathrm{Beta}(2,8)$. From Section 1.7 we know that the model we want is

$$\mathbb{P}[X = x, P \in B] = \int_B \mathbb{P}[M_p = x] f_P(p)\,dp \tag{2.2}$$

where $x \in \{0,\dots,10\}$ and $B \subseteq \mathbb{R}$. In particular, Lemma 1.7.1 tells us that $X|_{\{P=p\}} \overset{\mathrm{d}}{=} \mathrm{Bin}(10,p)$.

If we wanted to sample X, with the random parameter given by P, we can use a two-step procedure. First, we sample the value of p according to $P \sim \mathrm{Beta}(2,8)$. Then, we sample

©Nic Freeman, University of Sheffield, 2024.

$M_p \sim \text{Bin}(10, p)$, using whichever value of p we obtained. The result of this procedure is our model, X, for the data. We'll come back to this example shortly.

We are now ready to set up the set of type of models we are interested in. We will refer to them as *Bayesian models*, for a reason that will become clear in the next section. They come in two flavours: discrete data and continuous data. We will study the case of discrete data first, in the next section. In this case the model will be a generalized version of (2.2), where we allow any family of discrete random variables in place of (M_p) and any continuous random variable in place of P. We will handle the case of continuous data later on, in Chapter 3.

©Nic Freeman, University of Sheffield, 2024.

2.2 Discrete Bayesian models

We need two ingredients to construct the model:

1. Let $(M_\theta)_{\theta\in\Theta}$ be a family of discrete random variables with range $R \subseteq \mathbb{R}^n$ and parameter space $\Pi \subseteq \mathbb{R}^d$.
2. Let $f_\Theta : \mathbb{R}^d \to [0,\infty)$ be a probability density function with range Π.

The family (M_θ) is often referred to in textbooks as 'the' model, but to avoid confusion we will use the term *model family* instead. The possible values of θ for this family are given by the set Π, known as *parameter space* of the model. From Definition 1.3.1, all elements M_θ of the model family have the same range, which we call the *range* of the model.

The function f_Θ is known as the *prior* or more precisely the *prior probability density function*, for reasons that will be explained shortly. The distribution with p.d.f. f_Θ is known as the prior distribution. This is the distribution that we use to sample a random parameter from.

Definition 2.2.1 The *discrete Bayesian model* associated to (M_θ) and f_Θ is the random variable $(X,\Theta) \in \mathbb{R}^n \times \mathbb{R}^d$ with distribution given by

$$\mathbb{P}[X = x, \Theta \in A] = \int_A \mathbb{P}[M_\theta = x] f_\Theta(\theta)\, d\theta. \tag{2.3}$$

The random variable (X,Θ) is neither discrete nor continuous. The X part is discrete, and the Θ part (as we will see below) is continuous. Equation (2.3) is (1.13) in the special cases of a discrete model family. Let us unpack our notation in (2.3) a bit:

- $\theta \in \Pi$ is a particular choice of parameter;
- Θ is a random variable with range Π, the 'random version' of the parameter;
- X is the data sampled by the model, and x is a sample of this data.

From Section 1.7 and Lemma 1.7.1 we know that Θ has p.d.f. f_Θ and that $X|_{\{\Theta=\theta\}} \sim M_\theta$ whenever f_Θ is continuous at θ. To find the (marginal) distribution of the data X we set $A = \mathbb{R}^d$ in (2.3), giving the probability mass function

$$\mathbb{P}[X = x] = \int_{\mathbb{R}^d} \mathbb{P}[M_\theta = x] f_\Theta(\theta)\, d\theta. \tag{2.4}$$

This is known as the *sampling distribution* of our Bayesian model.

Example 2.2.2 To fit Example 2.1.1 into this notation, take M_θ to be the Binomial family with 10 trials, that is $M_p = \mathrm{Bin}(10,p)$ where $p = \theta$ takes values in $\Pi = [0,1]$. The prior chosen in Example 2.1.1 was $P = \Theta \sim \mathrm{Beta}(2,8)$, which takes values in $[0,1]$, and the p.d.f. we wrote down in (2.1). From (2.3) we obtain a discrete Bayesian model (X,P) with distribution given by

$$\begin{aligned}\mathbb{P}[X = x, P \in A] &= 72\binom{10}{x}\int_{A\cap[0,1]} p^x(1-p)^{10-x} p(1-p)^7\, dp \\ &= 72\binom{10}{x}\int_{A\cap[0,1]} p^{1+x}(1-p)^{17-x}\, dp.\end{aligned} \tag{2.5}$$

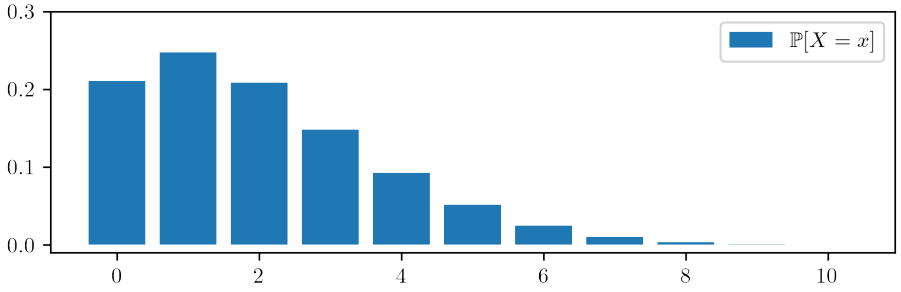
Equation (2.5) is precisely (2.2) with the binomial p.m.f. and beta p.d.f. from (2.1) filled in.

©Nic Freeman, University of Sheffield, 2024.

Putting $A = \mathbb{R}^d$ in (2.5) gives the distribution of X, also known as the sampling distribution of our model:

$$\mathbb{P}[X = x] = 72\binom{10}{x}\int_0^1 p^{1+x}(1-p)^{17-x}\,dp. \tag{2.6}$$

Equations (2.5) and (2.6) are not easy formulae. For the moment we will have to tolerate this sort of thing, before we think of some ways to make our calculations easier in Chapter 4. We can sketch $\mathbb{P}[X = x]$ numerically:



©Nic Freeman, University of Sheffield, 2024.

2.3 The posterior distribution

There is one more important piece of terminology associated to (X, Θ).

Definition 2.3.1 Let (X, Θ) be a discrete Bayesian model. For x such that $\mathbb{P}[X = x] > 0$, the random variable $\Theta|_{\{X=x\}}$ is known as the *posterior* of the model, and its distribution is known as the posterior distribution.

It will take a bit of work to understand why $\Theta|_{\{X=x\}}$ is important. Recall the prior Θ is a random variable. The distribution of Θ, which we usually specify via its p.d.f. $f_\Theta(\theta)$, is chosen based on our initial beliefs about where the true value of the unknown parameter θ might be. We then obtain some data x, coming from reality. The key idea is that $\Theta|_{\{X=x\}}$ will often be a better choice for the distribution of θ than our original choice Θ was. This is because $\Theta|_{\{X=x\}}$ incorporates information from the data that we have. We can summarise this idea as:

$$(\text{model parameters})|_{\{\text{model = the data we have}\}} = (\text{better model parameters}). \tag{2.7}$$

Here $|_{\{\dots\}}$ denotes conditioning. We will use the theory that we developed in Chapter 1 to make sense of (2.7).

The 'update' of Θ to $\Theta|_{\{X=x\}}$ is known as a *Bayesian update*, and the whole process is known as *Bayesian learning*. The terms 'prior' and 'posterior' are loose synonyms for 'before' and 'after', which is why they are used in Bayesian models.

All elements of the discrete family (M_θ) have the same range R, so $\mathbb{P}[M_\theta = x] > 0$ for all $x \in R$. It follows that the range of $\Theta|_{\{X=x\}}$ is the same as the range of Θ. Checking carefully against the two ingredients at the top of Section 2.2, this means that we can form a new discrete Bayesian model, with the same family (M_θ), the same parameter space Π, the same range R but with a new prior given by the probability density function $f_{\Theta|_{\{X=x\}}}$.

With our updated prior, we might want to use our new model to sample data (for whatever purpose). For this we use the sampling p.d.f. (2.4), but now applied to the model with prior $f_{\Theta|_{\{X=x\}}}$. There is a special piece of terminology for this:

Definition 2.3.2 Let (X, Θ) be a discrete Bayesian model, with parameter space Π and model family $(M_\theta)_{\theta\in\Pi}$. Let x be an item of data. The *predictive distribution* is the distribution of X' where

$$\mathbb{P}[X' = x'] = \int_{\mathbb{R}^d} \mathbb{P}[M_\theta = x'] f_{\Theta|_{\{X=x\}}}(\theta)\, d\theta. \tag{2.8}$$

Here, X' has the same range as X.

©Nic Freeman, University of Sheffield, 2024.

Example 2.3.3 Let us take the model from Example 2.1.1. We noted in Example 2.2.2 that this model was a discrete Bayesian model. Recall that our model was intended to model the number of successes from a set of 10 independent but identical experiments, where each experiment has an unknown probability p of success.

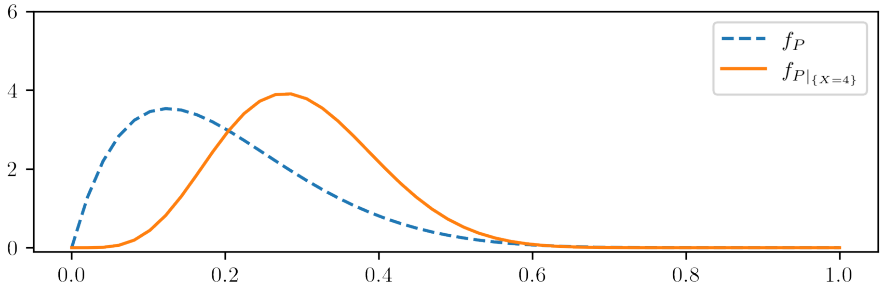
The model family we used is $M_p = \text{Bin}(10, p)$, with only one parameter p, so we take $\theta = p$. The range of this family is $R = \{0, 1, \dots, 10\}$. The parameter space is $\Pi = [0, 1]$, and our prior was $\Theta = P \sim \text{Beta}(2, 8)$, which has p.d.f. $f_P(p)$ that we sketched in Example 2.1.1.

Suppose that we learn that a scientist has carried out the 10 experiments, and obtained $x = 4$ successes. We can use Lemma 1.5.1 to compute the posterior $P|_{\{X=4\}}$, as follows. For $B \subseteq \mathbb{R}$,

$$\begin{aligned}
\mathbb{P}\left[P|_{\{X=4\}} \in B\right] &= \frac{\mathbb{P}[X = 4, P \in B]}{\mathbb{P}[X = 4]} \\
&= \frac{72\binom{10}{4}\int_{B\cap[0,1]} p^{1+4}(1-p)^{17-4}\,dp}{72\binom{10}{4}\int_{[0,1]} p^{1+4}(1-p)^{17-4}\,dp} \qquad (2.9) \\
&= \frac{\int_{B\cap[0,1]} p^5(1-p)^{13}\,dp}{\mathcal{B}(6,14)} \\
&= \int_{B\cap[0,1]} \frac{1}{\mathcal{B}(6,14)} p^5(1-p)^{13}\,dp \\
&= \int_B f_{\text{Beta}(6,14)}(p)\,dp \qquad (2.10)
\end{aligned}$$

where $\mathcal{B}(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx$ is the Beta function, which gives the normalizing constant of the Beta(α, β) distribution. To deduce (2.9) we have used (2.5) for the top and (2.6) for the bottom, with $x = 4$. Note that a factor of $72\binom{10}{4}$ cancels on the top and bottom.

From (2.10) we can recognize $P|_{\{X=4\}} \sim \text{Beta}(6, 14)$ as the posterior distribution of (X, P), given the data $x = 4$. Plotting the density functions of P and $P|_{\{X=4\}}$ gives the following:

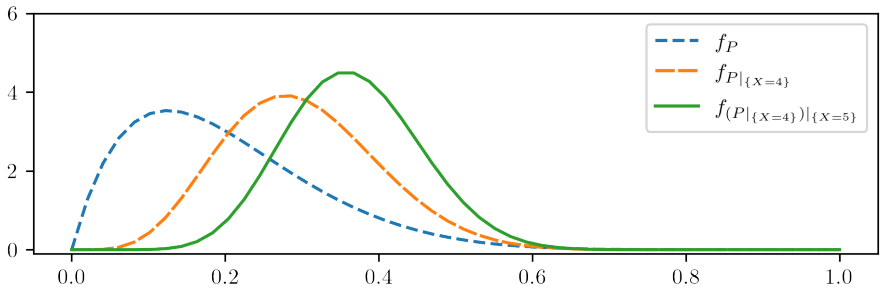


The effect of incorporating the datapoint $x = 4$ is visible here. Updating our prior p.d.f. f_P to the posterior p.d.f. $f_{P|_{\{X=4\}}}$ has resulting in the mass (i.e. the area underneath the curve) moving rightwards, towards the value $p = 0.4$ that corresponds to having 4/10 successful experiments. The p.d.f. $f_{P|_{\{X=4\}}}$ feels both the influence of the prior f_P, as well as the Bayesian update from our data.

Suppose that we are now given a second piece of data, which is that a second set of 10 experiments was done, with $x = 5$ successes. We can use the posterior Beta$(6, 14)$ that we obtained before as our new prior distribution, to incorporate our improved knowledge about the parameter θ. We keep the rest of our model as before, and do another Bayesian update to find a

©Nic Freeman, University of Sheffield, 2024.

new posterior distribution. Going through the same calculations (we'll omit the details) it turns out that the new posterior is Beta(11, 19). Including this into our plot we obtain:



We've labelled our new posterior as $f_{(P|_{\{X=4\}})|_{\{X=5\}}}$, to reflect the fact that we've done two updates (this is rather lazy notation!). We can see the influence of the new data: our second data point $x = 5$ again corresponds to a higher rate of success than our (updated) model anticipated, which again pulls the mass of the distribution rightwards.

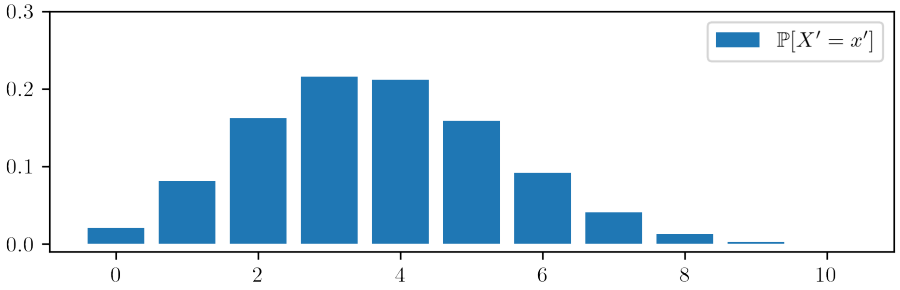
Our model still feels the effect of our initial prior, in the sense that the mean of $B(11, 19)$ is $11/30 \approx 0.37$, which corresponds to a lower success rate than both of our data points. We only have two data points, so it is perhaps best that the initial guess we were given has not been forgotten.

We can also see that the distributions are becoming less spread out with each update. This reflects our models becoming more confident (of the posterior distribution they suggest for θ) as we feed them more data. More precisely we can measure the standard deviations becoming smaller: from the reference sheet in Appendix A, Beta(α, β) has variance $\frac{\alpha\beta}{(\alpha+\beta)^2(1+\alpha+\beta)}$, from which the sequence of standard deviations comes out as 0.12, 0.10 and 0.087, each rounded to two significant figures.

Putting our final posterior Beta(11, 19) back into the same model family, we obtain from (2.8) that the predictive distribution given by our analysis is

$$\begin{aligned}
\mathbb{P}[X' = x'] &= \int_0^1 f_{\mathrm{Bin}(10,p)}(x') f_{\mathrm{Beta}(11,19)}(p)\, dp \\
&= \int_0^1 \binom{10}{x'} p^{x'}(1-p)^{10-x'} \frac{1}{\mathcal{B}(11,19)} p^{10}(1-p)^{18}\, dp \\
&= \frac{1}{\mathcal{B}(11,19)} \binom{10}{x'} \int_0^1 p^{10+x'}(1-p)^{28-x'}\, dp.
\end{aligned}$$

for $x' = 0, 1, \ldots, 10$. Plotting this p.m.f. gives:



©Nic Freeman, University of Sheffield, 2024.

We can compare this figure to the sampling p.d.f. in Example 2.2.2, from before we did any Bayesian updates. As we would expect, our estimate of the number of successful experiments now places more weight on larger values.

Calculating the distribution of $P|_{\{X=4\}}$ in Example 2.3.3 was a bit complicated. For practical purposes we will need to find easier ways of doing Bayesian updates. We'll do that in Chapter 4, but first we'll need to finish our development of the theory.

Once you become comfortable with Bayesian updates, it will be tempting to try and compare this new method of statistical inference to things you already know, and to wonder 'which is better?'. We will think about this in Chapter 6, but there is no simple answer.

In Example 2.3.3 we've used several sketches to help us understand the distributions we came across. Exercise **2.1** provides you with template code for doing so yourself. You will find that code useful for several other exercises too.

©Nic Freeman, University of Sheffield, 2024.

2.4 Bayesian updates

We now write down a general version of the method in Example 2.3.3, in the form of a theorem that we can apply once per update step. It is worth noting that in many practical situations only one update step is actually needed. This would normally be the case if we receive all the relevant data at the same time. If our data arrives gradually (e.g. once per year from an annual survey) then we can provide on-going analysis by carrying out an update step whenever new data arrives.

In Example 2.3.3 we only had one parameter, so our parameter space was $\Pi \subseteq \mathbb{R}$, and we only had one piece of data (per update) so our model for the data was a random variable X taking values in $\mathbb{R}$. In general we will want some number $d \in \mathbb{N}$ of parameters, so we take $\Pi \subseteq \mathbb{R}^d$, and we will want to handle some number $n \in \mathbb{N}$ of datapoints at once, so we let X take values in $\mathbb{R}^n$.

Theorem 2.4.1 (Bayesian updates for discrete data) *Let (X, Θ) be a discrete Bayesian model, with parameter space Π, prior p.d.f. f_Θ, model family (M_θ) and range R. Suppose that $x \in R$. Then the posterior $\Theta|_{\{X=x\}}$ is a continuous random variable and has p.d.f.*

$$f_{\Theta|_{\{X=x\}}}(\theta) = \frac{1}{Z}\mathbb{P}[M_\theta = x] f_\Theta(\theta) \tag{2.11}$$

where $Z = \int_\Pi \mathbb{P}[M_\theta = x] f_\Theta(\theta)\, d\theta$. The range of $\Theta|_{\{X=x\}}$ is Π, the same range as for Θ.

Proof of Theorem 2.4.1: Let (X, Θ) be a discrete Bayesian model as given and let $x \in R$. Note that $\mathbb{P}[M_\theta = x] > 0$ for all θ because (M_θ) is a discrete model family, so from (2.4) we have that $\mathbb{P}[X = x] > 0$. By Lemma 1.5.1

$$\mathbb{P}\left[\Theta|_{\{X=x\}} \in B\right] = \frac{\mathbb{P}[X = x, \Theta \in B]}{\mathbb{P}[X = x]} = \frac{\int_B \mathbb{P}[M_\theta = x] f_\Theta(\theta)\, d\theta}{\int_{\mathbb{R}^d} \mathbb{P}[M_\theta = x] f_\Theta(\theta)\, d\theta} = \int_B \frac{1}{Z}\mathbb{P}[M_\theta = x] f_\Theta(\theta)\, d\theta.$$

where $Z = \int_{\mathbb{R}^d} \mathbb{P}[M_\theta = x] f_\Theta(\theta)\, d\theta$. The denominator here comes from (2.4) and the numerator from (2.3).

The definition of a discrete Bayesian model gives that the prior Θ has range Π, so we may assume $f_\Theta(\theta) = 0$ for $\theta \notin \Pi$. Hence $Z = \int_\Pi \mathbb{P}[M_\theta = x] f_\Theta(\theta)\, d\theta$. It follows that $\Theta|_{\{X=x\}}$ is a continuous random variable with p.d.f. as in (2.11).

Also, $f_{\Theta|_{\{X=x\}}}(\theta) > 0$ if and only if $\theta \in \Pi$, so $\Theta|_{\{X=x\}}$ also has range Π. ■

Now that we have Theorem 2.4.1, and in particular equation (2.11), we should use it to calculate posterior densities – instead of falling back on Definition 1.4.2. For example, in Example 2.3.3 we can go straight from our description of the model to writing down the p.d.f. of $P|_{\{X=4\}}$ as

$$\begin{aligned} f_{P|_{\{X=4\}}}(p) &= \frac{1}{Z}\mathbb{P}[\mathrm{Bin}(10, p) = 4] f_{\mathrm{Beta}(2,8)}(p) \\ &= \frac{\binom{10}{4}\mathcal{B}(2,8)}{Z}\, p^4(1-p)^{10-4} p^{2-1}(1-p)^{8-1} \\ &= \frac{1}{Z'}\, p^5(1-p)^{13} \end{aligned} \tag{2.12}$$

for $p \in [0, 1]$, and zero elsewhere. Note that we have written $\frac{1}{Z'} = \frac{\binom{10}{4}\mathcal{B}(2,8)}{Z}$ in the last line, without having to do any computation. We only need to care about the part of the formula that depends on p, because the rest will be a normalizing constant. From Lemma 1.2.5 we can recognize (2.12) as the p.d.f. of the Beta$(6, 14)$ distribution.

©Nic Freeman, University of Sheffield, 2024.

2.4.1 Historical notes (⊘)

Equation (2.11) is often known as Bayes' rule, after Thomas Bayes (1701-1761). Bayes was one of the first mathematicians to study conditional probability, although he only became interested by it in later life and did not publish his work. Instead, it was edited and published by Richard Price (1723-1791) after Bayes' death. Both Bayes and Price were primarily interested in philosophy – statistics barely existed at the time and mathematics had only recently discovered calculus. In fact, what Bayes discovered is much closer to Lemma 1.4.1 in the special case of discrete random variables.

The concept of Bayesian inference first appears in work of Pierre-Simon Laplace (1749-1827). It was originally known as 'inverse probability' and kept this name up until the 1950s, during which the term 'Bayesian' became used instead. This makes Bayesian methods one of the oldest parts of statistics. By comparison, techniques based on maximum likelihood estimators (MLEs) were not introduced until the 1920s.

During the middle part of the 20th century, statistics was dominated by techniques based on MLEs and Bayesian techniques fell out of fashion. They become popular again with the development of modern computing power in the 1980s and 1990s, when it was realized (as we will see in Chapter 8) that Bayesian updates could be performed numerically without relying on families of well-known distributions. This provided the possibility of writing down highly complex Bayesian models whilst still having them 'learn' from data.

©Nic Freeman, University of Sheffield, 2024.

2.5 Exercises on Chapter 2

You can find formulae for named distributions in Appendix A.

⋆ **2.1** *This exercise provides template code for drawing several sketches of distributions, which you will find helpful in many later exercises.*

Use a computer package of your choice to complete the following questions. You will need the file `2_dist_sketching.ipynb` if you use Python, or `2_dist_sketching.Rmd` if you use R.

(a) You will find code that produces a sketch of the probability density functions of the Exp(3) and Exp(5) distributions. Modify this code to produce a sketch of the probability density functions of the Gamma$(4, 5)$ and Gamma$(6, 7)$ distributions.

(b) You will find code that produces a sketch of the Geometric$(\frac{1}{2})$ distribution. Modify this code to produce a sketch of the Bin$(10, \frac{2}{3})$ distribution.

(c) In this question we look at distributions in the form of equation (2.4).
You will find code that produces a sketch of the discrete distribution with p.m.f.

$$\mathbb{P}[X = x] = \int_0^1 \mathbb{P}[\text{Bin}(10, p) = x] f_{\text{Beta}(2,3)}(p)\, dp$$

for $x \in \{0, 1, \ldots, 10\}$. Note that this distribution is the sampling distribution associated to a discrete Bayesian model with model family Bin$(10, p)$ and prior $P \sim$ Beta$(2, 3)$.
Modify this code to produce a sketch of the discrete distribution with p.m.f.

$$\mathbb{P}[X = x] = \int_0^1 \mathbb{P}[\text{Geometric}(p) = x] f_{\text{Beta}(\frac{1}{2},\frac{1}{2})}(p)\, dp$$

for $x \in \{0, 1, \ldots, \}$.

⋆⋆ **2.2** Let (X, P) be a discrete Bayesian model with model family $M_p \sim$ Geometric(p) where $p \in [0, 1]$.

We regard M_p as a model for the number of times an experiment fails before the experiment is successful. The probability of success on each try is $p \in [0, 1]$, which is an unknown parameter. We assume that the experiments are independent of each other.

(a) Write down the probability mass function $\mathbb{P}[M_p = n]$, and the range of this model.

(b) Take a prior $P \sim$ Uniform$([0, 1])$. Given the single data point $x = 5$, show that the posterior distribution is given by $P|_{\{X=x\}} \sim$ Beta$(2, 5)$.

(c) Use a computer package of your choice to sketch the p.d.f. of this distribution, alongside the prior distribution.

(d) A second Bayesian update is made using a new data point, $x = 9$. Find the new posterior distribution and add it to your sketch from (c).

(e) Write down the p.m.f. of the predictive distribution, after the second update. Use a computer package of your choice to sketch it.

⋆⋆ **2.3** Let (X, Λ) be a discrete Bayesian model with model family $M_\lambda \sim$ Poisson(λ), where $\lambda \in (0, \infty)$. Take the prior to be $\Lambda \sim$ Exp(5). Find the distribution of the posterior $\Lambda|_{\{x=7\}}$ and write down the p.m.f. of the predictive distribution.

©Nic Freeman, University of Sheffield, 2024.

Chapter 3

Continuous Bayesian models

In this chapter we expand our results from Chapter 2 to also cover continuous data. This has the consequence that we will make more use of probability density functions than in Chapter 2, which causes some of the formulae to change and/or simplify. The key ideas do not change.

©Nic Freeman, University of Sheffield, 2024.

3.1 Continuous Bayesian models

In this section we construct a version of the model from Section 2.2 that is suitable for continuous data. We use a continuous family of random variables (M_θ), in place of the discrete family used in Section 2.2. It will behave in much the same way, except that when we used the p.m.f. of the discrete random variable M_θ, we will now use the p.d.f. of the continuous random variable M_θ.

We need two ingredients to construct the model:

1. Let $(M_\theta)_{\theta\in\Theta}$ be a family of continuous variables with range $R \subseteq \mathbb{R}^n$ and parameter space $\Pi \subseteq \mathbb{R}^d$. Write $f_{M_\theta} : \mathbb{R}^n \to [0,\infty)$ for the p.d.f. of M_θ.
2. Let $f_\Theta : \mathbb{R}^d \to [0,\infty)$ be a probability density function with range Π.

We used the same terminology as in Section 2.2: we refer to the family (M_θ) as the *model family*, and we will also use this term for (f_{M_θ}). We say that Π is the parameter space of the model and R is the range of the model. The random variable Θ, with p.d.f. f_Θ, is known as the *prior* of the model.

Definition 3.1.1 The *continuous Bayesian model* associated to (M_θ) and f is the random variable $(X,\Theta) \in \mathbb{R}^n \times \mathbb{R}^d$ with distribution given by

$$\mathbb{P}[X \in A, \Theta \in B] = \int_B \mathbb{P}[M_\theta \in A] f_\Theta(\theta)\, d\theta = \int_B \int_A f_{M_\theta}(x) f_\Theta(\theta)\, dx\, d\theta. \tag{3.1}$$

The random variable (X,Θ) is continuous, with p.d.f. $f_{M_\theta}(x)f_\Theta(\theta)$.

The symbols θ, Θ, x, X and have the same interpretations as listed in Section 2.2, and we won't repeat that list here. A warning: note that the p.d.f. $f_{M_\theta}(x)f_\Theta(\theta)$ in (3.1) is *not* in a factorized form $g(x)h(\theta)$ because $f_{M_\theta}(x)$ depends on both θ and x. Just as in Section 2.2, in general X and Θ are dependent random variables.

As in the discrete case, from Section 1.7 and Lemma 1.7.1 we have that Θ has p.d.f. f_Θ, and that $X|_{\{\Theta=\theta\}} \overset{\mathrm{d}}{=} M_\theta$ whenever f_Θ is continuous at θ. To find the (marginal) p.d.f. of the data X we must instead integrate out the θ variable from the joint p.d.f., giving

$$f_X(x) = \int_{\mathbb{R}^d} f_{M_\theta}(x) f_\Theta(\theta)\, d\theta. \tag{3.2}$$

This is known as the *sampling p.d.f.* or sampling density of the model, and the distribution of X is the *sampling distribution*.

The *posterior* of the model is $\Theta|_{\{X=x\}}$, which is defined using Lemma 1.6.1. For the same reasons as in the discrete case, we can hope that using $\Theta|_{\{X=x\}}$ in place of Θ will result in an improved model. Let's work out the distribution of the posterior (in general) before we do an example with some data.

Theorem 3.1.2 (Bayesian updates for continuous data) *Let (X,Θ) be a continuous Bayesian model, with parameter space Π, prior p.d.f. f_Θ, model family (M_θ) and range R. Write f_{M_θ} for the p.d.f. of M_θ. Suppose that $x \in R$. Then the posterior $\Theta|_{\{X=x\}}$ is a continuous random variable and has p.d.f.*

$$f_{\Theta|_{\{X=x\}}}(\theta) = \frac{1}{Z} f_{M_\theta}(x) f_\Theta(\theta) \tag{3.3}$$

where $Z = \int_\Pi f_{M_\theta}(x) f_\Theta(\theta)\, d\theta$. The range of $\Theta|_{\{X=x\}}$ is Π, the same range as for Θ.

©Nic Freeman, University of Sheffield, 2024.

PROOF: The proof is similar to our proof of Theorem 2.4.1, except that we use Lemma 1.6.1 (in place of Lemma 1.5.1) to find the p.d.f. of $\Theta|_{\{X=x\}}$.

A difficulty is that Lemma 1.6.1 requires continuity conditions, which are not always satisfied in the situation here (although they are often are). For that reason, we will only give a proof covering the special case where both $f_{M_\theta}(x)$ and $f_\Theta(\theta)$ are continuous functions. In that case, from Lemma 1.6.1 we have that

$$f_{\Theta|_{\{X=x\}}}(\theta) = \frac{f_{(X,\Theta)}(x,\theta)}{f_X(x)} = \frac{f_{M_\theta}(x) f_\Theta(\theta)}{f_X(x)}. \tag{3.4}$$

We have used p.d.f. coming from Definition 3.1.1 in the numerator above. For the denominator, we already found $f_X(x)$ in (3.2), which gives $f_X(x) = Z$. This gives (3.3). The definition of a continuous Bayesian model requires that the prior Θ has range Π, so $f_\Theta(\theta) > 0$ if and only if $\theta \notin \Pi$. Since $x \in \mathbb{R}$ we have $f_{M_\theta}(x) > 0$ for all $\theta \in \Pi$. Hence the range of $\Theta|_{\{X=x\}}$ is Π. ■

Remark 3.1.3 (⊘) In fact equation (3.3) only holds for almost all $x \in R$, but it works for all x when we have 'enough continuity' in some suitable sense. This is generally sufficient for practical purposes and we won't worry about this issue within these notes. To see a natural case where (3.3) fails for a particular choice of x, take $\Theta \sim \Gamma(\frac{1}{4}, 1)$ and $M_\theta \sim N(0, \theta)$, with the data $x = 0$. Then according to (3.3) we have $f_{\Theta|_{\{X=0\}}}(\theta) = \frac{1}{Z}\frac{1}{\sqrt{2\pi\theta}} e^{-0^2/2\theta} \frac{1^{1/4}}{\Gamma(1/4)} \theta^{1/4-1} e^{-\theta} = \frac{1}{Z'}\theta^{-5/4} e^{-\theta}$ for $\theta > 0$. This does not define a p.d.f. since $\int_0^\infty \theta^{-5/4} e^{-\theta}\, d\theta = \infty$. The problem stems from fact that $(x, \theta) \mapsto f_{M_\theta}(x)$ is discontinuous at $(0, 0)$, which causes the continuity conditions mentioned in the above proof to fail. In this case for any $x \neq 0$ we do obtain a posterior p.d.f. that integrates to one.

Equation (3.3) is a 'continuous data' analogue of equation (2.11). Both equations are often known as (versions of) Bayes' rule, for the historical reasons that we discussed in Section 2.3. In both cases Z is a normalizating constant, ensuring that the p.d.f. of $\Theta|_{\{X=x\}}$ integrates to 1. The only difference between (2.11) and (3.3) is that:

- (2.11) features the p.m.f. $\mathbb{P}[M_\theta = x]$ of the (discrete) model family;
- (3.3) features the p.d.f. $f_{M_\theta}(x)$ of the (continuous) model family.

Lastly, once we have obtained $\Theta|_{\{X=x\}}$ we construct a new Bayesian model with $\Theta|_{\{X=x\}}$ in place of Θ. As before:

Definition 3.1.4 The *predictive distribution* is given by replacing the prior Θ with the posterior $\Theta|_{\{X=x\}}$, inside the sampling distribution. Hence, from (3.2), the predictive distribution is of a continuous random variable with p.d.f.

$$f_{X'}(x') = \int_{\mathbb{R}^d} f_{M_\theta}(x') f_{\Theta|_{\{X=x\}}}(\theta)\, d\theta. \tag{3.5}$$

©Nic Freeman, University of Sheffield, 2024.

3.2 Notation: independent data

We often want to construct a Bayesian model where the data corresponds to n independent, identically distribution samples from some common distribution. That is, we want our model family to be of the form $X = (X_1, \dots, X_n)$ where the X_i are independent with the same distribution. It is helpful to have some notation for this.

Given a pair of random variables Y and Z, we write $Y \otimes Z$ for the random variable (Y, Z) formed of a copy of Y and a copy of Z that are independent of each other. We will tend to use this notation in combination with named distributions. For example, $X \sim \mathrm{N}(0,1) \otimes \mathrm{N}(0,1)$ means that $X = (X_1, X_2)$ is a pair of independent $\mathrm{N}(0,1)$ random variables. When we want to create n copies we will use a superscript $\otimes n$, so $X \sim \mathrm{N}(0,1)^{\otimes n}$ means that $X = (X_1, \dots, X_n)$ is a sequence of n independent $\mathrm{N}(0,1)$ random variables.

Note that if X has range R then $X^{\otimes n}$ has range R^n.

Example 3.2.1 Recall that a Bernoulli trial is a random variable $X \sim \mathrm{Bernoulli}(p)$, with distribution $\mathbb{P}[X = 1] = p$ and $\mathbb{P}[X = 0] = 1 - p$. The standard relationship between Bernoulli trials and the Binomial distribution can be written as follows: if $(X_i)_{i=1}^n \sim \mathrm{Bernoulli}(p)^{\otimes n}$ then $\sum_{i=1}^n X_i \sim \mathrm{Bin}(n, p)$.

Example 3.2.2 Let M be a continuous random variable with p.d.f. f_M and let $X \overset{\mathrm{d}}{=} M^{\otimes n}$. Then X has p.d.f.

$$f_X(x) = \prod_{i=1}^n f_M(x_i),$$

where $x = (x_1, \dots, x_n)$. A similar relationship applies in the case of discrete random variables, to probability mass functions.

©Nic Freeman, University of Sheffield, 2024.

Example 3.2.3 We are interested to model the duration of time that people spend on social activities. We will use data from the 2015 American Time Use survey, corresponding to the category 'socializing and communicating with others'.

We decide to model the time spent on a single social activity as an exponential random variable $\text{Exp}(\lambda)$, where λ is an unknown parameter. This is a common model for time durations. Our data consists of $n = 50$ independent responses, each of which tells us the duration that was spent on a single social activity, in minutes. This gives us the model family

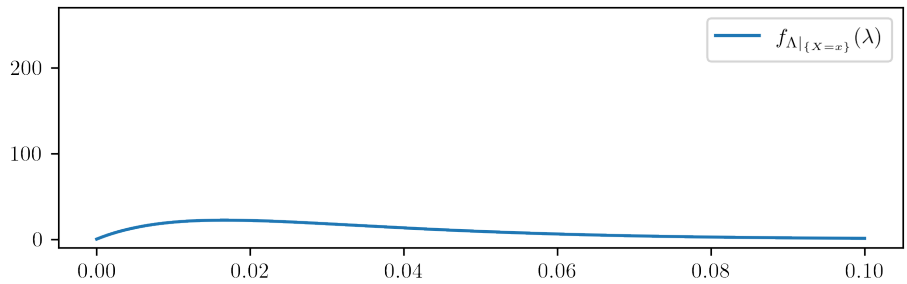
$$M_\lambda = \text{Exp}(\lambda)^{\otimes 50}$$

which has p.d.f.

$$f_{M_\lambda}(x) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^{50} e^{-\lambda \sum_1^{50} x_i}. \tag{3.6}$$

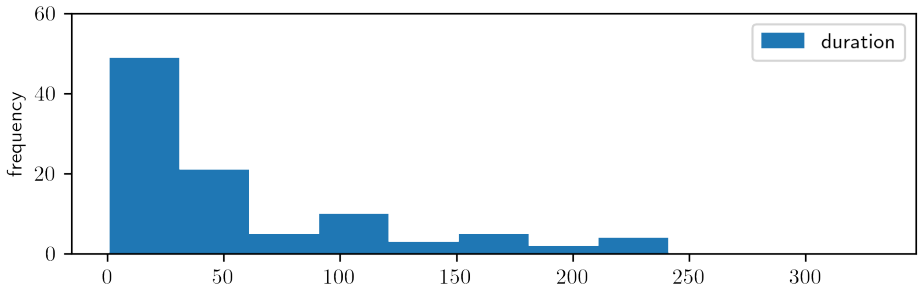
A single item of data has range $(0, \infty)$ so the range of our model is $(0, \infty)^{50}$.

We need to choose a prior for λ. As in Example 2.2.2 we will make a somewhat arbitrary choice, because for now our focus is on understanding how Bayesian updates work. Our prior for the duration of a social activity will be $\Lambda \sim \text{Gamma}(2, 60)$,



We can check that this prior sits roughly within the region of parameters that we would expect: it has $\mathbb{E}[\Lambda] = \frac{2}{60} = \frac{1}{30}$, which corresponds to an average social activity time of 30 minutes, because $\mathbb{E}[\text{Exp}(\lambda)] = \frac{1}{\lambda}$.

We represent our data as a vector $x = (x_1, \ldots, x_{50})$. A histogram of the data is as follows:



It satisfies $\sum_1^n x_i = 6638$, which we can fill into (3.6).

We can find the posterior distribution $\Lambda|_{\{X=x\}}$ using Theorem 3.1.2. It has p.d.f.

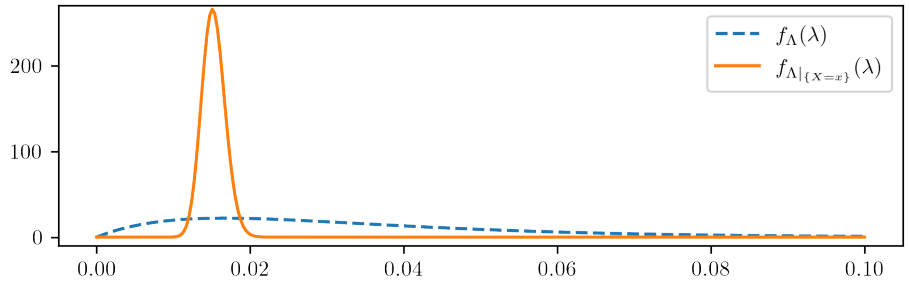
$$f_{\Lambda|_{\{X=x\}}}(\lambda) = \frac{1}{Z} f_{M_\lambda}(x) f_{\text{Gamma}(2,60)}(\lambda)$$

©Nic Freeman, University of Sheffield, 2024.

$$= \frac{1}{Z}\lambda^{50}e^{-6638\lambda}\frac{60^2}{\Gamma(2)}\lambda^{2-1}e^{-60\lambda}$$
$$= \frac{1}{Z'}\lambda^{51}e^{-6698\lambda} \tag{3.7}$$

Note that here we have absorbed the factor $\frac{60^2}{\Gamma(2)}$ into the normalizing constant $\frac{1}{Z}$ to obtain a new normalizing constant $\frac{1}{Z'}$. We know that (3.7) is a probability density function, so by Lemma 1.2.5 we have that $\Lambda|_{\{X=x\}} \sim \text{Gamma}(52, 6698)$, and we know that $\frac{1}{Z'}$ must be the normalizing constant of the $\Gamma(52, 6698)$ distribution.

Plotting the prior and posterior probability density functions together gives

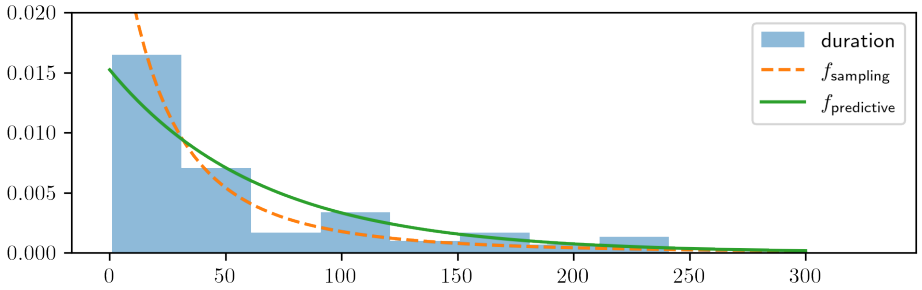


Here we see that, even though our prior is spread out across a fairly wide range of values, the posterior has focused very precisely on a small region. By comparison to Example 2.3.3, we have a lot more data here, so our analysis here has produced a higher level of confidence in the best choice of parameter values. Consequently our choice of prior has mattered less than it did in Example 2.3.3.

It is sensible to compare the results of our analysis our the histogram of the data x. Our model is for 50 independent samples, so technically the sampling and predictive distributions of our model generate 50 real-valued samples, which is awkward to sketch. Instead we use the sampling and predictive distributions for a single data point (i.e. the $n = 1$ case of our model). This gives sampling and predictive distributions, from (3.2) and (3.5), with probability density functions

$$f_{\text{sampling}}(x_1) = \int_{\mathbb{R}_d} f_{\text{Exp}(\lambda)}(x_1) f_{\Gamma(2,60)}(\lambda)\, d\lambda,$$
$$f_{\text{predictive}}(x_1) = \int_{\mathbb{R}_d} f_{\text{Exp}(\lambda)}(x_1') f_{\Gamma(52,6698)}(\lambda)\, d\lambda.$$

Comparing these to the data, we obtain



©Nic Freeman, University of Sheffield, 2024.

It is clear that the predictive distribution is a better match for the data than the sampling distribution. To make the comparison we have scaled the total area of the histogram to be 1, to match the fact the area under probability density functions is also 1.

Remark 3.2.4 In both Chapter 2 and 3 we have used continuous distributions for our random parameters. In principle we could use discrete distributions instead i.e. Π would become a finite set and Θ would only be allowed to take values in Π. We would need to slightly modify (2.11) and (3.3) for such cases. There aren't any families of common distributions where the parameters spaces are discrete, and in practice we rarely have a reason to want models of this type. We won't study models of this type within our course.

©Nic Freeman, University of Sheffield, 2024.

3.3 Exercises on Chapter 3

You can find formulae for named distributions in Appendix A.

⋆ **3.1** *This exercise continues Exercise 2.1. It provides template code for drawing several sketches of distributions, which you will find helpful in many later exercises.*

Inside the files `2_dist_sketching.ipynb` and `2_dist_sketching.Rmd`, below the parts corresponding to Exercise 2.1, you will find the code for sketching

$$f_X(x_1) = \int_{\mathbb{R}_d} f_{\text{Exp}(\lambda)}(x_1) f_{\text{Gamma}(2,60)}(\lambda)\, d\lambda,$$

which is the p.d.f. of the sampling distribution (for a single item of data) in Example 3.2.3.

(a) Modify this code to sketch the p.d.f. of the sampling distribution of the continuous Bayesian model (X, Θ) with model family $M_\theta = \Gamma(2, \theta)$ and prior $\Theta \sim \text{Exp}(1)$.

(b) Do the same as in (a), for the continuous Bayesian model (X, Θ) with model family $M_\theta = \text{N}(\theta, 1)$ and prior $\Theta \sim \text{N}(0, 1)$.

⋆⋆ **3.2** Let $M_\theta \sim \text{Exp}(\theta)$, where θ takes values in the parameter space $\Pi = (0, \infty)$. Let (X, Θ) be the Bayesian model with this model family and prior $\Theta \sim \text{Gamma}(2, 3)$.

(a) Given the single data point $x = 2$, show that the posterior $\Theta|_{\{X=2\}}$ has the Gamma$(3, 5)$ distribution.

(b) (i) Show that the sampling distribution of the model has p.d.f.

$$f_X(x) = \begin{cases} \frac{18}{(x+3)^3} & \text{for } x > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.8}$$

Hint: Use that $\int_0^\infty f_{\text{Gamma}(3,x+3)}(\theta)\, d\theta = 1$ *to help with the integral.*

(ii) Find the predictive distribution in similar form to (3.8).

(c) Now consider the model $M_\theta \sim (X_1, \ldots, X_n)$, where $n \in \mathbb{N}$ and the X_i are independent Exp(θ) random variables. Use the same prior $\Theta \sim \text{Gamma}(2, 3)$ and the data $x = (x_1, \ldots, x_n)$, where $x_i \in (0, \infty)$ for all $i = 1, \ldots, n$. Show that the posterior distribution has the Gamma$(n + 2,\, 3 + z)$ distribution, where $z = \sum_1^n x_i$.

⋆ **3.3** With a computer package of your choice, sketch the prior and posterior probability density functions from Exercise 3.2(a)/(b) on the same graph.

On a separate graph, use the explicit formula you found in Exercise 3.2(a)/(b) to sketch the sampling and predictive distributions. Modify your code from Exercise 3.1 to sketch the same functions, but without using your explicit formulae. Check that the results agree.

⋆ **3.4** (a) Look at the left hand column of the reference sheet 'Conditional Probability and Related Formulae' in Appendix A. For each item listed there, identify which Section, Lemma, equation, or other part of Chapter 1 it comes from.

(b) Do the same for the left hand column of the reference sheet 'Bayesian Models and Related Formulae' (excluding the last item), with Chapters 2 and 3.

©Nic Freeman, University of Sheffield, 2024.

3.5 Let $M_\theta \sim \text{Uniform}([0,\theta])$ be the continuous uniform distribution on $[0,\theta]$. Let (X,Θ) be a Bayesian model with model family $(M_\theta)_{\theta\in\Pi}$, with parameter space $\Pi = (0,\infty)$. Take the prior to be $\Theta \sim \text{Pareto}(3,1)$.

⋆⋆ (a) Suppose that we have the datapoint $x = \frac{1}{2}$. Show that the posterior $\Theta|_{\{X=\frac{1}{2}\}}$ has distribution Pareto$(4,1)$.

⋆⋆⋆ (b) Suppose instead that we had the data point $x = 5$. Find the posterior distribution $\Theta|_{\{X=5\}}$ and calculate the p.d.f. of the resulting predictive distribution.

⋆⋆⋆ **3.6** Let (X,Θ) be a continuous Bayesian model with parameter space Π. Suppose that $A \subseteq \Pi$ with $\mathbb{P}[\Theta \in A] > 0$. Show that $X|_{\{\Theta\in A\}}$ is a continuous random variable with p.d.f.

$$f_{X|_{\{\Theta\in A\}}}(x) = \int_A f_{M_\theta}(x) f_{\Theta|_{\{\Theta\in A\}}}(\theta)\, d\theta.$$

⋆⋆⋆ **3.7** Let (X,Θ) be a continuous Bayesian model, with range $R_X \subseteq \mathbb{R}$ and parameter space $\Pi \subseteq \mathbb{R}$, and with model family (M_θ). Let f_{M_θ} denote the p.d.f. of M_θ and let f_Θ denote the p.d.f. of Θ.

Consider a second continuous Bayesian model (X',Θ) with the same prior, the same range and parameter space, but with model family (M'_θ) given by

$$f_{M'_\theta}(x) = \int_{\mathbb{R}} f_{M_\theta}(x-y)\kappa(y)\, dy. \tag{3.9}$$

We require that $\kappa : \mathbb{R} \to [0,\infty)$ and $\int_{\mathbb{R}} \kappa(y)\, dy = 1$.

(a) Check that $\int_{\mathbb{R}} f_{M'_\theta}(x)\, dx = 1$.

(b) Show that the posterior density of (X',Θ) satisfies

$$f_{\Theta|_{\{X'=x\}}}(\theta) \propto \int_{\mathbb{R}} f_{\Theta|_{\{X=x-y\}}}(\theta)\kappa(y)\, dy. \tag{3.10}$$

(c) The operation in (3.9) is known as the *convolution* of f_{M_θ} with κ, and the function κ is known as the *kernel* of the convolution.

Consider the case $\kappa(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$. Investigate the connection between convolutions and sums of random variables. Use what you discover to write down (in words) a heuristic interpretation of the connection between Model 1 and Model 2, and also of equation (3.10).

©Nic Freeman, University of Sheffield, 2024.

Chapter 4

Conjugate priors

Recall that we define a Bayesian model (X, Θ) using a prior Θ and a model family $(M_\theta)_{\theta\in\Pi}$. The main result from Chapters 2 and 3 is that we can (at least, in principle) obtain the distribution of the posterior $\Theta|_{\{X=x\}}$. If the prior and posterior are from the same family of distributions then we say that this family of distributions is *conjugate* to the model family.

Note that the setup here involves *two* families of distributions: (1) the model family (M_θ) and (2) the *conjugate* family, to which the prior and posterior both belong. The formal definition is a bit of a mouthful:

Definition 4.0.1 Let $(M_\theta)_{\theta\in\Pi}$ and $(T_a)_{a\in A}$ be two model families, with parameter spaces Π and A respectively. We say that (M_θ) and (T_a) are a *conjugate pair* if whenever (X, Θ) is a Bayesian model with model family (M_θ) and prior $\Theta \sim T_a$, for all $x \in R_X$ there exists $b \in A$ such that $\Theta|_{\{X=x\}} \sim T_b$. We say that the family (T_a) is a *conjugate prior* for (M_θ).

The point of using conjugate pairs is that, to specify a Bayesian update step, we only need to describe how the parameters of the prior distribution should change, to obtain the posterior distribution. This is in general much simpler than (2.11) and (3.3). We will describe a few conjugate pairs in this chapter and discuss their limitations in Section 4.6.

©Nic Freeman, University of Sheffield, 2024.

4.1 Notation: proportionality

In (2.12) and (3.7) we used $\frac{1}{Z}$ and $\frac{1}{Z'}$ for normalizing constants. It was helpful not to worry about exactly what the value of these constant were. In longer calculations we might need to use several different normalizing constants in this way, and it is helpful to have some notation for doing so (beyond simply $\frac{1}{Z''}$, $\frac{1}{Z'''}$ and so on).

Definition 4.1.1 Let f and g be functions within the same domain. We write $f \propto g$ if there exists $C \in (0, \infty)$ such that $f(x) = Cg(x)$ for all x. In words, f is said to be *proportional* to g.

The relation $\propto$ has several nice properties, which are easy to check and are left for you in Exercise 4.9. For example, for any function f we have $f \propto f$. Also, $f \propto g$ if and only if $g \propto f$ and, lastly, if $f \propto g$ and $g \propto h$ then $f \propto h$. We'll use these properties frequently in calculations, without further comment.

Example 4.1.2 Using the notation $\propto$, Lemma 1.2.5 says that for random variables X and Y:

- If X and Y are discrete and $p_X \propto p_Y$ then $X \overset{\mathrm{d}}{=} Y$.
- If X and Y are continuous and $f_X \propto f_Y$ then $X \overset{\mathrm{d}}{=} Y$.

We will often use Lemma 1.2.5 in this way from now on, including in the next example.

Example 4.1.3 The calculation in (2.12) can be written simply as

$$\begin{aligned} f_{P|_{\{X=4\}}}(p) &\propto \mathbb{P}[\mathrm{Bin}(10,p) = 4] f_{\mathrm{Beta}(2,8)}(p) \\ &\propto p^4(1-p)^{10-4} p^{2-1}(1-p)^{8-1} \\ &\propto p^5(1-p)^{13}. \end{aligned}$$

It follows immediately from Lemma 1.2.5 that $P|_{\{X=4\}} \sim \mathrm{Beta}(6, 14)$.

Example 4.1.4 More generally, the key equations (2.11) and (3.3) from Theorems 2.4.1 and 3.1.2 can be written

$$\begin{aligned} f_{\Theta|_{\{X=x\}}}(\theta) &\propto p_{M_\theta}(x) f_\Theta(\theta) && \text{for discrete Bayesian models,} \\ f_{\Theta|_{\{X=x\}}}(\theta) &\propto f_{M_\theta}(x) f_\Theta(\theta) && \text{for continuous Bayesian models.} \end{aligned}$$

We will often use Theorems 2.4.1 and 3.1.2 in this way from now on.

A complication of using $\propto$ is that the symbol does not explicitly specify which variables should be treated as part of the proportionality, and which other variables can be treated as constants. For our purposes there is a simple way to resolve this difficulty. We we use $\propto$ there will, in most cases, be a function on the left of the first $\propto$ that appears within a calculation. The arguments of that function (not including subscripts) are the variables that proportionality applies to; everything else can be treated as constant in so far as $\propto$ is concerned.

©Nic Freeman, University of Sheffield, 2024.

4.1.1 The Beta-Binomial pair

Here is our first example of a conjugate pair, which generalizes the calculations in Example 2.1.1-2.3.3.

Lemma 4.1.5 (Beta-Bernoulli conjugate pair) *Let $n \in \mathbb{N}$. Let (X, Θ) be a discrete Bayesian model with model family $M_\theta \sim \text{Bernoulli}(\theta)^{\otimes n}$ and parameter $\theta \in [0,1]$. Suppose that the prior is $\Theta \sim \text{Beta}(a,b)$ and let $x \in \{0,1\}^n$. Then the posterior is $\Theta|_{\{X=x\}} \sim \text{Beta}(a+k, b+n-k)$ where $k = \sum_1^n x_i$.*

Proof: Note that k is the number of Bernoulli trials that generate a 1, and that we have n trials in total. Under M_p, each trial has probability p of generating 1. From Theorem 2.4.1 we have that for $\theta \in [0,1]$

$$\begin{aligned} f_{\Theta|_{\{X=x\}}}(\theta) &\propto \mathbb{P}[\text{Bernoulli}(\theta)^{\otimes n} = x] f_{\text{Beta}(a,b)}(\theta) \\ &\propto \left(\prod_{i=1}^n \mathbb{P}[\text{Bernoulli}(\theta) = x_i] \right) f_{\text{Beta}(a,b)}(\theta) \\ &\propto \left(\theta^k (1-\theta)^{n-k} \right) \left(\frac{1}{\mathcal{B}(a,b)} \theta^{a-1} (1-\theta)^{b-1} \right) \\ &\propto \theta^{a+k-1} (1-\theta)^{b+n-k-1}. \end{aligned}$$

By Lemma 1.2.5 we recognize this p.d.f. as $\Theta|_{\{X=x\}} \sim \text{Beta}(a+k, b+n-k)$. ■

The value of k has an intuitive interpretation, because it is the number of successful trials observed in our data x (here we take a trial to result in 1 if successful, and 0 if failed). Looking back at Example 2.3.3, this allows us to do all of the Bayesian update calculations with one easy piece of arithmetic.

More generally, we can use the Binomial distribution in place of the Bernoulli distribution, as in the next lemma.

Lemma 4.1.6 (Beta-Binomial conjugate pair) *Let $n, m_i \in \mathbb{N}$. Let (X, Θ) be a discrete Bayesian model with model family*

$$M_\theta \sim \text{Bin}(m_1, \theta) \otimes \ldots \otimes \text{Bin}(m_n, \theta).$$

with the parameter $\theta \in \Pi = [0,1]$. Suppose that the prior is $\Theta \sim \text{Beta}(a,b)$ and let $x = (x_1, \ldots, x_n)$ where $x_i \in \{0, \ldots, m_i\}$. Then the posterior is $\Theta|_{\{X=x\}} \sim \text{Beta}\left(a + \sum_1^n x_i, b + \sum_1^n m_i - \sum_1^n x_i\right)$.

Proof: From Theorem 2.4.1 we have that for $\theta \in [0,1]$

$$\begin{aligned} f_{\Theta|_{\{X=x\}}}(\theta) &\propto \left(\prod_{i=1}^n \binom{m_i}{x_i} \theta^{x_i} (1-\theta)^{m_i - x_i} \right) \left(\frac{1}{\mathcal{B}(a,b)} \theta^{a-1} (1-\theta)^{b-1} \right) \\ &\propto \theta^{a + \sum_1^n x_i - 1} (1-\theta)^{b + \sum_1^n m_i - \sum_1^n x_i - 1}. \end{aligned}$$

By Lemma 1.2.5 we recognize this p.d.f. as $\Theta|_{\{X=x\}} \sim \text{Beta}\left(a + \sum_1^n x_i, b + \sum_1^n m_i - \sum_1^n x_i\right)$. ■

Remark 4.1.7 (⊘) There is a further generalization of this model to experiments that can have many possible outcomes. It involves the Dirichlet and multinomial distributions. It is not much more complicated, but we won't cover it within this course.

©Nic Freeman, University of Sheffield, 2024.

4.2 Two more examples of conjugate pairs

There are several examples of conjugate pairs in the exercises at the end of this chapter. We include a couple more here, the first of which generalizes the calculations in Example 3.2.3.

Lemma 4.2.1 (Gamma-Exponential conjugate pair) *Let $n \in \mathbb{N}$. Let (X, Λ) be a continuous Bayesian model with model family $M_\lambda \sim \mathrm{Exp}(\lambda)^{\otimes n}$ and parameter $\lambda \in (0, \infty)$. Suppose that the prior is $\Lambda \sim \mathrm{Gamma}(\alpha, \beta)$ and let $x \in (0, \infty)^n$. Then the posterior is $\Lambda|_{\{X=x\}} \sim \mathrm{Gamma}(\alpha + n, \beta + \sum_1^n x_i)$.*

Proof: From Theorem 3.1.2 we have that for $\lambda \in (0, \infty)$

$$\begin{aligned} f_{\Lambda|_{\{X=x\}}}(\lambda) &\propto f_{\mathrm{Exp}(\lambda)^{\otimes n}}(x) f_{\mathrm{Gamma}(\alpha,\beta)}(\lambda) \\ &\propto \left(\prod_{i=1}^n \lambda e^{-\lambda x_i} \right) \left(\frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \right) \\ &\propto \lambda^n e^{-\lambda \sum_1^n x_i} \lambda^{\alpha-1} e^{-\beta\lambda}. \\ &\propto \lambda^{\alpha+n-1} e^{-\lambda(\beta + \sum_1^n x_i)}, \end{aligned}$$

By Lemma 1.2.5 we recognize this p.d.f. as $\Theta|_{\{X=x\}} \sim \mathrm{Gamma}(\alpha + n, \beta + \sum_1^n x_i)$. ■

We'll now do a more complicated example in which the constant of proportionality would change multiple times – if we were to write it in, which we won't, thanks to $\propto$. In Section 4.5 we will look at Bayesian inference for the normal distribution where both μ and σ are unknown parameters. For now we view σ as fixed, so the mean μ is the only parameter.

Lemma 4.2.2 (Normal-Normal conjugate pair) *Let $u \in \mathbb{R}$ and $\sigma, s > 0$. Let (X, Θ) be a continuous Bayesian model with model family $M_\theta \sim \mathrm{N}(\theta, \sigma^2)^{\otimes n}$ and parameter $\theta \in \mathbb{R}$. Suppose that the prior is $\Theta \sim \mathrm{N}(u, s^2)$ and let $x \in \mathbb{R}^n$. Then*

$$\Theta|_{\{X=x\}} \sim \mathrm{N}\left(\frac{\frac{1}{\sigma^2}\sum_1^n x_i + \frac{u}{s^2}}{\frac{n}{\sigma^2} + \frac{1}{s^2}}, \ \frac{1}{\frac{n}{\sigma^2} + \frac{1}{s^2}} \right). \tag{4.1}$$

Proof: From Theorem 3.1.2 we have that for $\theta \in \mathbb{R}$

$$\begin{aligned} f_{\Theta|_{\{X=x\}}}(\theta) &\propto f_{\mathrm{N}(\theta,\sigma^2)^{\otimes n}}(x) f_{\mathrm{N}(u,s^2)}(\theta) \\ &\propto \left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2\sigma^2}} \right) \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(\theta-u)^2}{2s^2}} \right) \\ &\propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2 - \frac{1}{2s^2}(\theta - u)^2 \right) \\ &\propto \exp\big(-\mathcal{Q}(\theta) \big) \end{aligned}$$

where

$$\mathcal{Q}(\theta) = \theta^2 \overbrace{\left(\frac{n}{2\sigma^2} + \frac{1}{2s^2} \right)}^{A} - 2\theta \overbrace{\left(\frac{1}{2\sigma^2} \sum_{i=1}^n x_i + \frac{u}{2s^2} \right)}^{B} + \overbrace{\left(\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2 + \frac{u^2}{2s^2} \right)}^{C}.$$

©Nic Freeman, University of Sheffield, 2024.

Completing the square in $\mathcal{Q}(\theta)$, using the general form of completing the square (which you can find on the reference sheet in Appendix A), we have that

$$f_{\Theta|_{\{X=x\}}}(\theta) \propto \exp\left(-A\left(\theta - \frac{B}{A}\right)^2 + C - \frac{B^2}{A}\right)$$
$$\propto \exp\left(-\frac{1}{2(\frac{1}{2A})}\left(\theta - \tfrac{B}{A}\right)^2\right).$$

By Lemma 1.2.5 we recognize this p.d.f. as $\Theta|_{\{X=x\}} \sim \mathrm{N}\left(\frac{B}{A}, \frac{1}{2A}\right)$. We have

$$\frac{B}{A} = \frac{\frac{1}{\sigma^2}\sum_{i=1}^n x_i + \frac{u}{s^2}}{\frac{n}{\sigma^2} + \frac{1}{s^2}} \qquad \text{and} \qquad \frac{1}{2A} = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{s^2}},$$

as required. Note that in the first term we have cancelled a factor of $\frac{1}{2}$ from both $\mathcal{A}$ and $\mathcal{B}$. ■

From (4.1) we can see that the variance will decrease as $n \to \infty$, and that for large n it will be $\approx \frac{\sigma^2}{n}$. This agrees with our experience in Example 4.2.4, in which case we had $\sigma^2 = s^2 = 0.4^2$, giving variance $\frac{0.4^2}{n}$. Recall that in our discussion at the end of Example 4.2.4 we noted that the posterior variance had become very small after only 10 obserations, despite the prior having a reasonably large variance.

In the formulae we obtained in (4.1) each time a variance appears, for both σ^2 and s^2, it appears on the bottom of a fraction. This suggests that we might obtain nicer formulae if we instead to parameterize the normal distribution as $\mathrm{N}(\mu, \frac{1}{\tau})$, where by by comparison to our usual parametrization we have written $\tau = \frac{1}{\sigma^2}$. It is common to do this in Bayesian statistics and the variable τ is then known as *precision*. We will do this, for example, in Exercise **4.4** which considers the Normal distribution with fixed mean and unknown variance.

You can find a table of conjugate pairs on the reference sheets in Appendix A, below the tables of named distributions. It contains all of the examples within this chapter; there is no need for you to memorize the formulae.

Remark 4.2.3 You are now ready to start on all of the exercises for this chapter.

©Nic Freeman, University of Sheffield, 2024.

Example 4.2.4 Speed cameras are used to measure the speed of individual cars. They do so by recording two images of a moving car, with the second image being captured a fixed time after the first image. By analysing the two images the camera can tell how far the car has travelled in that time, which gives an estimate of its speed. This is not an easy process and there is some degree of error involved.

Suppose that we are trying to assess if the manufacturers description of the error is accurate. The manufacturer claims that, if the true speed is 30 (miles per hour) then the speed recorded by the camera can be modelled as a $\mathrm{N}(30, 0.16)$ random variable.

We construct an experiment to test this. We set up a camera and drive 10 cars past it, each travelling at 30 miles per hour (let us assume this can be done accurately, which is not unrealistic using modern cruise control). The camera records speeds of

$$30.9, \quad 29.9, \quad 30.1, \quad 30.3, \quad 29.7, \quad 30.1, \quad 30.1, \quad 29.2, \quad 30.6, \quad 30.4. \tag{4.2}$$

We record this data as $x = (x_i)_{i=1}^{10}$.

We will come back to this example in future but for now let us assume, for simplicity, that we believe the manufacturers claim that the data will have a normal distribution with variance 0.16. We want to see if mean matches up with the figure claimed. We'll use the model family

$$M_\theta \sim N(\theta, 0.16)^{\otimes 10} \stackrel{\mathrm{d}}{=} N(\theta, 0.4^2)^{\otimes 10},$$

where the mean θ is an unknown parameter. The parameter space of this model is $\Pi = \mathbb{R}$ and the range of the model is $\mathbb{R}^{10}$.

For our prior we will use $\Theta \sim \mathrm{N}(30, 5^2)$ which has p.d.f.

$$f_\Theta(\theta) = \frac{1}{5\sqrt{2\pi}} e^{-(x-30)^2/10}.$$

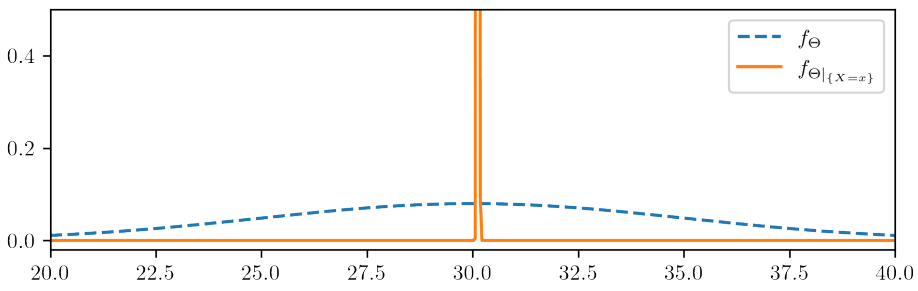
We will study techniques for choosing the prior in Chapter 5. For now our motivation is that we expect the true value for θ is about 30, but we don't have a lot of confidence in that, so we pick a fairly large value for the variance.

Remark 4.2.5 It is always sensible to think about what property of reality your 'true' parameter value represents. In this case, the true value of θ is the average speed that would be recorded by the camera, for a car that was travelling at exactly 30mph. We don't know this value.

By Lemma 4.2.2 the posterior distribution is

$$\Theta|_{\{X=x\}} \sim \mathrm{N}\left(\frac{\frac{1}{0.16}\sum_1^{10} x_i + \frac{30}{5^2}}{\frac{10}{0.16} + \frac{1}{5^2}}, \frac{1}{\frac{10}{0.16} + \frac{1}{5^2}} \right) \stackrel{\mathrm{d}}{=} \mathrm{N}\left(30.13, 0.04^2\right).$$

Here we fill in $\sum_1^{10} x_i = 301.4$ and round the parameters to two decimal places. As in our previous examples, let us compare the prior Θ to the posterior $\Theta|_{\{X=x\}}$.



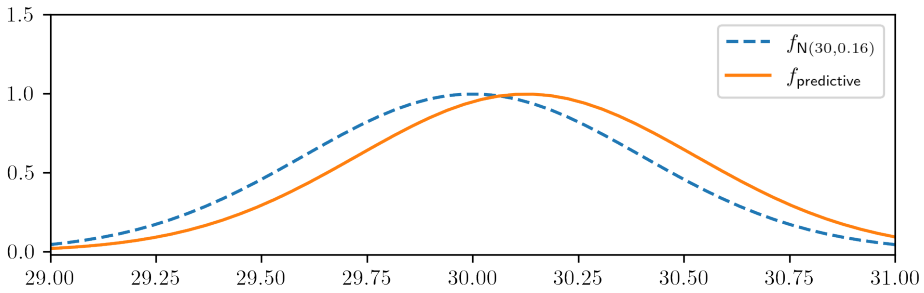
©Nic Freeman, University of Sheffield, 2024.

It is difficult to show them on the same axis, so we have had to miss out the top part of the curve $f_{\Theta|_{\{X=x\}}}$. The outcome is similar to Example 3.2.3, in that the posterior has focused in on a small region. Given our data (4.2) this seems sensible. The influence of the prior has largely been forgotten.

We were originally interested to compare the behaviour the manufacturer claimed that the camera would have, with the results of our experiment. To do so we should compare $\mathrm{N}(30, 0.16)$, which is what the manufacturer claimed our experiment should observe, with the predictive distribution from our data analysis. As in Example 3.2.3, what we want here is the predictive distribution for a single data point (i.e. the case $n = 1$). For that, our model family is $N(\theta, 0.4^2)$, and our posterior distribution for the unknown parameter θ is $\mathrm{N}(30.13, 0.04^2)$, which gives the p.d.f. of the predictive distribution for a single datapoint as

$$f_{\text{predictive}}(x) = \int_{\mathbb{R}} f_{\mathrm{N}(\theta,0.4^2)}(x) f_{\mathrm{N}(30.13,0.04^2)}(\theta)\, d\theta$$

which we can evaluate numerically[1]. We obtain



They are quite similar. If our predictive distribution is a true reflection of the cameras behaviour, it suggests that the camera may be overestimating speeds by a small amount. We would need to do some statistical testing before saying anything more, based on the 10 datapoints that we have, and we'll have to wait until Chapter 7 for that.

We'll return to this data again in Example 4.5.3, where will also treat the variance as an unknown parameter.

[1] In fact, some further calculation would reveal that this is also the p.d.f. of a normal distribution.

©Nic Freeman, University of Sheffield, 2024.

4.3 Conjugate pairs and the exponential family (⊘)

For continuous Bayesian models it is known that conjugate priors can only be found when the model family has a particular form. We'll restrict here to continuous models with one unknown parameter, but a similar picture holds in general. In particular, it must be from the *exponential family* of distributions. This term does not refer to the exponential distribution, but to a much wider class. We say that a real valued random variable Y is from the exponential family of distributions with parameter θ if it has a p.d.f. in the form

$$f_Y(y) = h(y)g(\theta)\exp\left(\theta T(y)\right). \tag{4.3}$$

Here h, g and T are arbitrary functions, with the restriction $h \geq 0$. Many of the families of distributions that you are familiar with can be fitted into this mould, including the normal, exponential, gamma, chi-squared, and beta distributions.

Take a Bayesian model (X, Θ) with model family given by (4.3), where both X and Θ take values in $\mathbb{R}$. That is,

$$f_{M_\theta}(x) = h(x)g(\theta)\exp\left(\theta T(x)\right) \tag{4.4}$$

for $x, \theta \in \mathbb{R}$. We'll focus on the version of this model that takes n independent items of data, in which case instead

$$f_{M_\theta}(x) = \left(\prod_{i=1}^{n} h(x_i)\right) g(\theta)^n \exp\left(\theta \sum_{i=1}^{n} T(x_i)\right). \tag{4.5}$$

The prior distribution that provides a conjugate pair to this model is given by

$$f_\Theta(\theta) \propto g(\theta)^a \exp\left(b\theta\right), \tag{4.6}$$

where $a > 0$ and $b \in \mathbb{R}$ are parameters Note that this distribution is a specialized version of the form in (4.3), and that the function g must be the same as in (4.5). Theorem 3.1.2 allows us to compute the posterior density

$$\begin{aligned} f_{\Theta|_{\{X=x\}}}(\theta) &\propto f_{M_\theta}(x) f_\Theta(\theta) \\ &\propto g(\theta)^{n+a} \exp\left[\theta\left(b + \sum_{i=1}^{n} T(x_i)\right)\right]. \end{aligned} \tag{4.7}$$

Therefore the Bayesian update in this case, from the parameters in (4.6) to (4.7) is that

$$(a, b) \mapsto \left(a + n,\, b + \sum_{i=1}^{n} T(x_i)\right).$$

Example 4.3.1 Taking $h(x) = 1$, $g(\theta) = \theta$ and $T(x) = x$ obtains the $\text{Exp}(\theta)$ in (4.4). Making the same choice for g, along with $a = \alpha$ and $b = -\beta$, obtains that $\text{Gamma}(\alpha, \beta)$ distribution in (4.6). Putting these choices for g and T into (4.7) and applying Lemma 1.2.5 gives that the posterior is $\text{Gamma}(\alpha + n, \beta + \sum_1^n x_i)$, as we already knew from Lemma 4.2.1.

Remark 4.3.2 There is also a version of this framework for multiple parameters, in which θ and b are row vectors and T is a column vector, and multiplication of these quantities is done via the dot product. We won't write down the details of that case here. There is also a version that applies to discrete Bayesian models, but we won't detail that either.

©Nic Freeman, University of Sheffield, 2024.

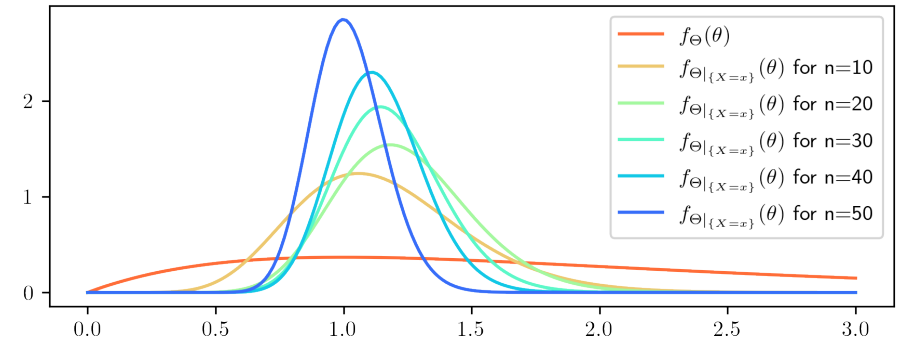
4.4 What if?

In this section we do some numerical experiments to illustrate the sort of things that go wrong if, for some reason, our model family or our prior is too unrealistic. We will first describe a situation where the inference works as intended, and then we'll do some things to break it.

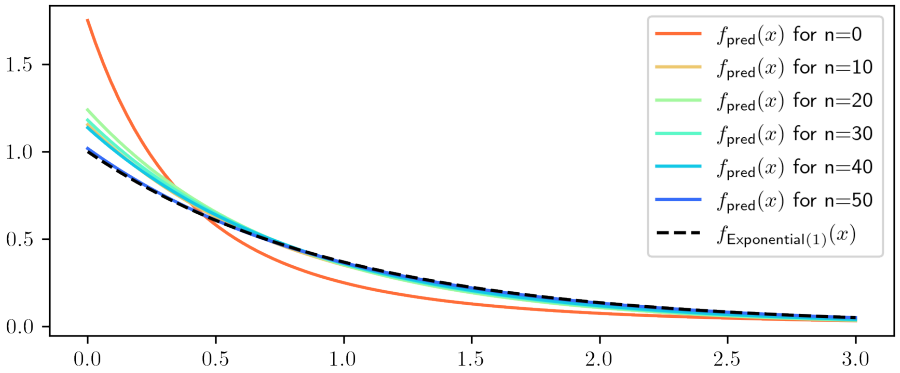
Example 4.4.1 Let (X, Θ) be a continuous Bayesian model with model family $M_\lambda \sim \text{Exp}(\lambda)^{\otimes n}$ and parameter $\lambda \in (0, \infty)$. Take the prior to be $\Theta \sim \Gamma(\alpha, \beta)$, where $\alpha, \beta \in (0, \infty)$.

For now, we choose the prior with $\alpha = \beta = 1$. We will feed our model data consisting of i.i.d. samples from the Exp(1) distribution. This corresponds to a true value of the parameter given by $\lambda^* = 1$. In other words, if we set $\Theta = \lambda$ in our Bayesian model then the samples it would produce for X would have exactly the same distribution as the data. So, we hope that Bayesian updates based on this data will result in a posterior density with its mass near to 1.

The posterior densities that resulted from various amounts of data x, sampled from $\text{Exp}(1)^{\otimes n}$, are as follows.



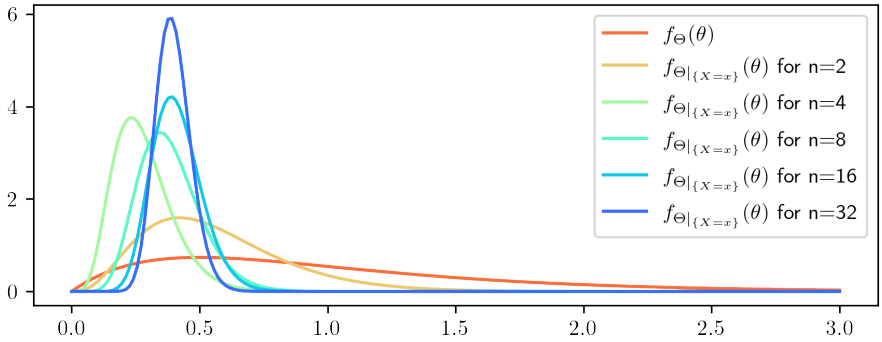
The Bayesian updates here were calculated using Lemma 4.2.1. As expected, although the mass of our prior is not close to the true value, we can see the posterior densities becoming more and more focused on the value 1. We can also see the corresponding predictive distributions (for a single datapoint) converging towards Exp(1). The density functions look like:



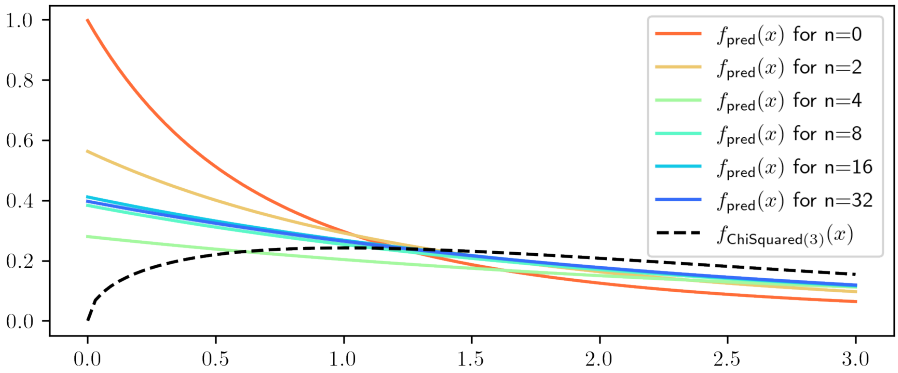
©Nic Freeman, University of Sheffield, 2024.

What if the model is wrong?

Example 4.4.2 Here we use the same model as in Example 4.4.1, but now we feed our model data consisting of i.i.d. samples from the ChiSquared(2) distribution. In this case there is no value of the parameter λ for which our model (X, Θ) is a good representation of the data. We are interested to see what happens:



As before, it looks like our model is trying to focus in on one particular value for λ. This time it looks to be homing in on a value a little below 0.5, although it not yet clear which. Note that the convergence appearing here actually seems faster than in Example 4.4.1 i.e. the values shown here for n are a bit smaller. Let us draw the probability density functions of the predictive distributions that come from the posteriors above, and compare them to the true density function from which our data was sampled:



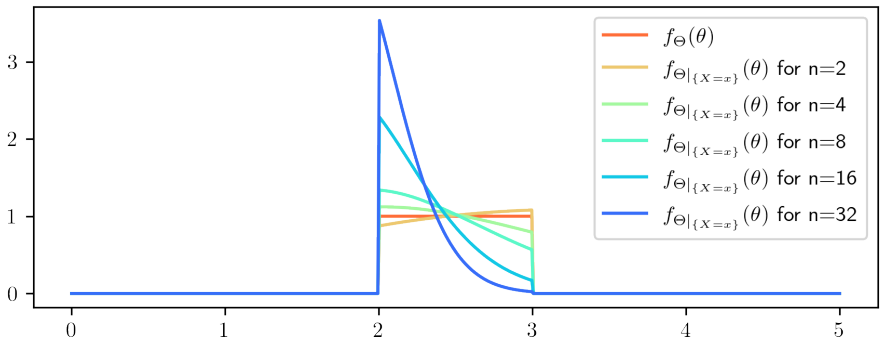
The density functions of the predictive distributions also seem to be converging as n gets larger, as we would expect from the convergence suggested in the previous graph. But what they are converging towards is not much like the p.d.f. from which the data was actually generated (shown as the black dotted line). It puts far too much weight on small values, for example. If we hadn't looked at these graphs we might not have noticed that anything was wrong.

In other situations, where we choose a model that is not able to represent the data well, the posterior distributions won't converge at all as $n \to \infty$. In such cases the predictive distributions tend to be wildly wrong by comparison to the data.

©Nic Freeman, University of Sheffield, 2024.

What if the prior excludes the true value?

Example 4.4.3 We'll use the model Example 4.4.1 again, but we'll change the prior, to be $\Theta \sim \text{Uniform}([2,3])$, the continuous uniform distribution with range $[2,3]$. We'll feed the model data consisting of i.i.d. samples from the Exp(1) distribution, corresponding (as in Example 4.4.1) to a true value $\lambda^* = 1$ for the parameter. The key point is that in this case the true value is outside of the range of the prior. Let's examine what happens now:



The posteriors are focusing on the value $\lambda = 2$, but the most conspicuous feature is that the posterior distributions place no weight at all outside of $[2,3]$. In fact we knew this in advance, because Theorems 2.4.1 and 3.1.2 both include the fact that the range of $\Theta|_{\{X=x\}}$ is equal to the range of Θ, in this case $[2,3]$. This means that, no matter how much data we give to our model, it will never converge towards the true value $\lambda^* = 1$. It is trying to do the next best thing, and get as close as possible. We won't draw the predictive distributions for this case, but the same story applies there.

There is an important message to take away from this example. When we assign zero prior probability to some region of the parameter space, our model interprets it as an instruction that we do not *ever* want to consider parameters in that region, even if it later turns out that the data fits that region better. If we chose to do this based on our own (or our expert colleagues) opinions, and it turns out that we are wrong, then we have made a serious error.

To avoid that situation, it is generally agreed that the range of the prior should include all values of the parameter that are physically possible, with a view to the situation that we are trying to model. This is known as *Cromwell's rule*, based on the quotation

> *I beseech you, in the bowels of Christ, think it possible that you may be mistaken!*

which Oliver Cromwell famously wrote in a letter to the council of the Church of Scotland in 1650, in an attempt to persuade them not to support an invasion of England. Cromwell failed to persuade them and shortly afterwards Scotland did invade England. Cromwell's own forces decisively won the resulting battle for the English side.

We have followed Cromwell's rule in Example 2.3.3, where the parameter p represented a probability and our prior had range $[0,1]$. Similarly, in Example 4.2.4 the parameter θ represented a speed and our prior had range $\mathbb{R}$. In the context of that example we would hope that negative values of the parameter would not be plausible (or the speed camera is seriously malfunctioning!) but we still allowed, in our prior, a small probability that this might occur.

©Nic Freeman, University of Sheffield, 2024.

4.5 The normal distribution with unknown mean and variance

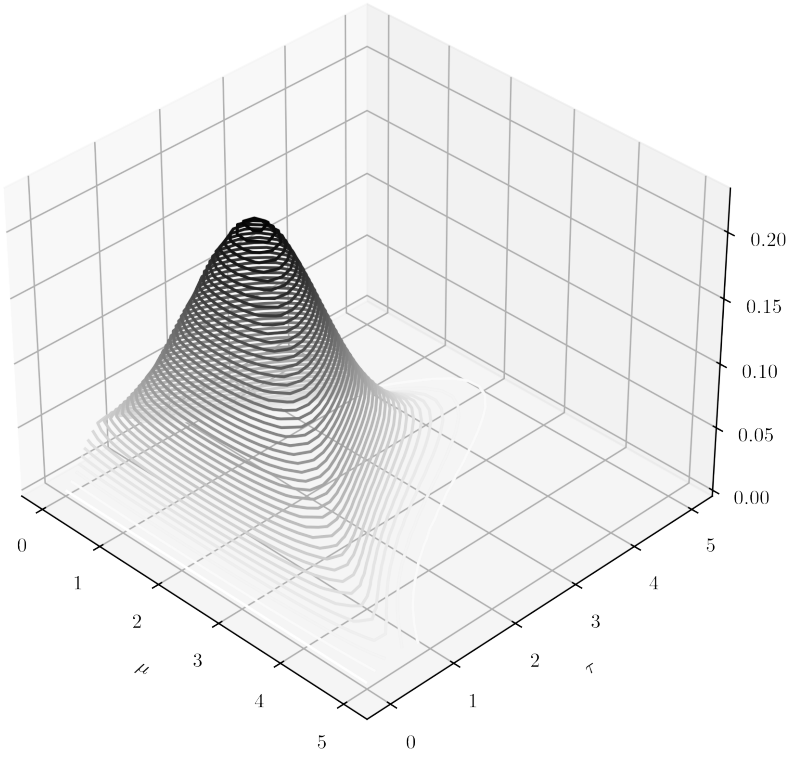
We've considered the normal distribution with a fixed variance, but with unknown mean, in Lemma 4.2.2 and Example 4.2.4. The situation of fixed mean and unknown variance is treated in Exercise 4.4. We'll deal here with the case where both the mean and variance are unknown parameters.

In the formulae obtained in Lemma 4.2.2, both variables relating the variance (σ^2 and s^2) only appeared as $\frac{1}{\sigma^2}$ and $\frac{1}{s^2}$. This suggests that we would obtain neater formulae if we parameterized the variance part as $\tau = \frac{1}{\sigma^2}$. The parameter τ is known as *precision*. We will use this form in the next lemma, and also in Exercise 4.4.

The conjugate prior in this case is complicated. It is the Normal-Gamma distribution, written $\text{NGamma}(m, p, a, b)$ with p.d.f.

$$\begin{aligned} f_{\text{NGamma}(m,p,a,b)}(\mu,\tau) &= f_{\text{N}(m,\frac{1}{p\tau})}(\mu)\, f_{\text{Gamma}(a,b)}(\tau) && (4.8)\\ &= \sqrt{\frac{p\tau}{2\pi}} \exp\left(-\frac{p\tau}{2}(\mu-m)^2\right) \ \frac{b^a}{\Gamma(a)}\tau^{a-1}e^{-b\tau} \\ &\propto \tau^{a-\frac{1}{2}} \exp\left(-\frac{p\tau}{2}(\mu-m)^2 - b\tau\right). && (4.9) \end{aligned}$$

You can find this p.d.f. on the reference sheets in Appendix A. Note that it is a two-dimensional distribution, which we will use to construct random versions of the parameters μ and τ in $\text{N}(\mu, \frac{1}{\tau})$. The restrictions on the NGamma parameters are that $m \in \mathbb{R}$ and $p, a, b > 0$. The range is $\mu \in \mathbb{R}$ and $\tau > 0$. Here's a contour plot of the p.d.f. of $\text{NGamma}(2, 1, 3, 2)$ to give some idea of what is going on here:



©Nic Freeman, University of Sheffield, 2024.

Let us note a couple of facts about the NGamma(m, p, a, b) distribution before we proceed further. If $(U, T) \sim$ NGamma(m, p, a, b) then:

- The marginal distribution of T is Gamma(a, b).

  This follows easily from (4.8). Integrating out μ will remove the term $f_{\mathrm{N}(m, \frac{1}{p\tau})}(\mu)$, which is a p.d.f. and integrates to 1, leaving only term $f_{\mathrm{Gamma}(a,b)}(\tau)$.

- The conditional distribution of $U|_{\{T=\tau\}}$ is $\mathrm{N}(m, \frac{1}{p\tau})$.

  This also follows from (4.8), by applying Lemma 1.6.1. We already know that $f_T(t) = f_{\mathrm{Gamma}(a,b)}(\tau)$, so

$$f_{U|_{\{T=\tau\}}}(\mu) = \frac{f_{\mathrm{N}(m, \frac{1}{p\tau})}(\mu)\, f_T(\tau)}{f_T(\tau)} = f_{\mathrm{N}(m, \frac{1}{p\tau})}(\mu).$$

Note that we are using U as a capital μ and T as a capital τ, to preserve our usual relationship between random variables and the arguments of their probability density functions. These two facts won't be used in our proof of conjugacy, but hopefully they help explain the formula (4.8) and the picture below it.

Our next goal is to state the conjugacy between NGamma and the Normal distribution. We will need the *sample-mean-variance* identity, which states that for all $x \in \mathbb{R}^n$

$$\sum_{i=1}^{n}(x_i - \mu)^2 = ns^2 + n(\bar{x} - \mu)^2 \tag{4.10}$$

where $\bar{x} = \frac{1}{n}\sum_1^n x_i$ and $s^2 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2$.

Remark 4.5.1 (⊘) To deduce (4.10), let Z be a random variable with the uniform distribution on $\{x_1, \ldots, x_n\}$. Note that $\mathbb{E}[Z] = \bar{x}$ and $\mathrm{var}(Z) = s^2$. The identity follows from the fact that $\mathrm{var}(Z) = \mathrm{var}(Z - \mu) = \mathbb{E}[(Z-\mu)^2] - \mathbb{E}[Z - \mu]^2$.

Lemma 4.5.2 (Normal-NGamma conjugate pair) *Let (X, Θ) be a continuous Bayesian model with model family $M_{\mu,\tau} = \mathrm{N}(\mu, \frac{1}{\tau})^{\otimes n}$, with parameters $\mu \in \mathbb{R}$ and $\tau > 0$. Suppose that the prior is $\Theta = (U, T) \sim$ NGamma(m, p, a, b) and let $x \in \mathbb{R}^n$. Then $\Theta|_{\{X=x\}} \sim$ NGamma (m^*, p^*, a^*, b^*) where*

$$m^* = \frac{n\bar{x} + mp}{n+p} \qquad p^* = n + p$$
$$a^* = a + \frac{n}{2} \qquad b^* = b + \frac{n}{2}\left(s^2 + \frac{p}{n+p}(\bar{x} - m)^2\right),$$

where $\bar{x} = \frac{1}{n}\sum_1^n x_i$ and $s^2 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2$.

Proof: (⊘) From Theorem 3.1.2 we have that for all $\mu \in \mathbb{R}$ and $\tau > 0$,

$$\begin{aligned} f_{(U,T)|_{\{X=x\}}}(\mu, \tau) &\propto \left(\prod_{i=1}^{n}\sqrt{\frac{\tau}{2\pi}}\exp\left(-\frac{\tau(\mu - x_i)^2}{2}\right)\right)\tau^{a-\frac{1}{2}}\exp\left(-\frac{p\tau}{2}(\mu - m)^2 - b\tau\right) \\ &\propto \tau^{\frac{n}{2}+a-\frac{1}{2}}\exp\left[-\frac{\tau}{2}\left(\sum_{i=1}^{n}(\mu - x_i)^2 + p(\mu - m)^2 + 2b\right)\right] \\ &\propto \tau^{\frac{n}{2}+a-\frac{1}{2}}\exp\left[-\frac{\tau}{2}\left(ns^2 + n(\mu - \bar{x})^2 + p(\mu - m)^2 + 2b\right)\right] \end{aligned}$$

©Nic Freeman, University of Sheffield, 2024.

$$\propto \tau^{\frac{n}{2}+a-\frac{1}{2}} \exp\left[-\frac{\tau}{2}\mathcal{Q}(\mu)\right]$$

To deduce the third line we have used (4.10). We have

$$\mathcal{Q}(\mu) = \mu^2 \overbrace{(n+p)}^{A} - 2\mu \overbrace{(n\bar{x}+mp)}^{B} + \overbrace{\left(ns^2+n\bar{x}^2+pm^2+2b\right)}^{C}.$$

Completing the square (with the help of the reference sheet) we obtain $\mathcal{Q}(\mu) = A(\mu - \frac{B}{A})^2 + C - \frac{B^2}{A}$ and hence

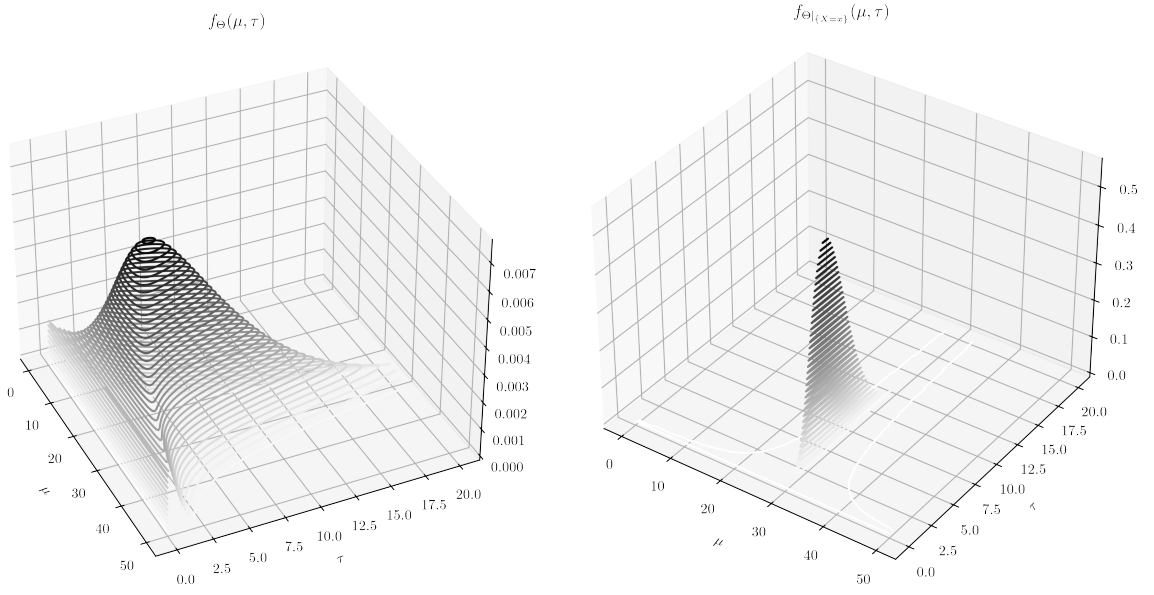
$$f_{(U,T)|\{X=x\}}(\mu,\tau) \propto \tau^{\frac{n}{2}+a-\frac{1}{2}} \exp\left[-\frac{\tau}{2}A\left(\mu-\frac{B}{A}\right)^2 - \tau\frac{1}{2}\left(C-\frac{B^2}{A}\right)\right].$$

The right hand side is above in the form of (4.9) and we can now extract the posterior NGamma parameters. From the exponent of τ we have $a^* = a + \frac{n}{2}$. From the term $(\mu - \ldots)^2$ we have $m^* = \frac{B}{A} = \frac{n\bar{x}+mp}{n+p}$ and from the coefficient of this term we have $p^* = A = n+p$. This leaves only

$$\begin{aligned}
b^* &= \frac{1}{2}\left(C - \frac{B^2}{A}\right) \\
&= \frac{1}{2}\left[ns^2 + n\bar{x}^2 + pm^2 + 2b - \frac{1}{n+p}(n\bar{x}+mp)^2\right] \\
&= b + \frac{ns^2}{2} + \frac{1}{n+p}\left[(n+p)n\bar{x}^2 + (n+p)pm^2 - n^2\bar{x}^2 + 2mpn\bar{x} - m^2p^2\right] \\
&= b + \frac{ns^2}{2} + \frac{1}{n+p}\left[pn\bar{x}^2 + npm^2 + 2mpn\bar{x}\right] \\
&= b + \frac{ns^2}{2} + \frac{np}{n+p}(\bar{x}-m)^2,
\end{aligned}$$

which completes the proof. ■

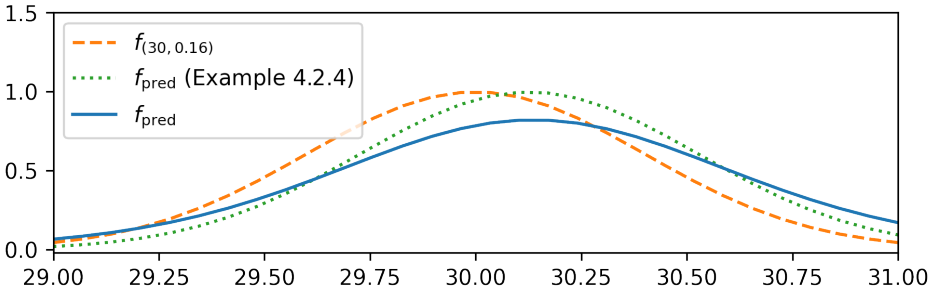
Example 4.5.3 Coming back to Example 4.2.4, regarding testing a speed camera, we can now do a Bayesian update where both the mean and variance are unknown parameters.



©Nic Freeman, University of Sheffield, 2024.

On the left we've taken an NGamma$(30, \frac{1}{10^2}, 1, \frac{1}{5})$ prior. Note that $p = \frac{1}{10^2}$ corresponds to standard deviation $= 10$, so our prior is well spread out about its mean $m = \mu = 30$ on the μ-axis. The values chosen for a and b ensure that the prior is also well spread out on the τ-axis. On the right we have the resulting NGamma$(30.14, 10.01, 6.00, 1.24)$ posterior. These posterior parameters were computing using Lemma 4.5.2 and with the same ten datapoints as in Example 4.2.4. As we would expect from Example 4.2.4, the mass of the posterior has focused close to $\mu \approx 30$. The parameter τ has focused on a wide range of fairly large values, but remember that $\tau = \frac{1}{\sigma^2}$ so the range of likely values for the standard deviation σ will in fact be a small range of small numbers.

Comparing the predictive densities gives:



Our new predictive distribution is still broadly similar to the manufacturers N$(30, 0.16)$ claim, but it now looks more spread out and the mean remains slightly higher than the manufacturers claim. A point that might worry us is that our new predictive p.d.f. is not close to zero at 31mph, whereas the manufacturer claims that it should be; our data suggests that the camera is more likely to overestimate speeds than the manufacturer has claimed. We can't reasonably say more without statistical testing, which we'll study in Chapter 7.

Think for a moment about how much numerical work has been done to produce a graph of the predictive pdf here. According to (3.5), for each point x' on the graph, to obtain $f_{\text{pred}}(x')$ we need to integrate $f_{\text{N}(\mu, \frac{1}{\tau})}(x') f_{\text{NGamma}(30.14,10.01,6.00,1.24)}(\mu, \tau)$ with respect to both μ and τ, over a region of $\mathbb{R}^2$ that includes the spike visible on the posterior density. Numerical integration over $\mathbb{R}^2$ is expensive – doing it once is not very noticeable, but doing it repeatedly usually is. The graph was made using 30 x-axis values and took 255 seconds to create, using `scipy.integrate.dblquad` for the integration. If we had three unknown parameters then we would have to integrate in $\mathbb{R}^3$, which is even worse. In fact, problems of this type with several parameters quickly become computationally infeasible via direct numerical integration. They need a more subtle numerical technique, which we'll introduce in Chapter 8.

©Nic Freeman, University of Sheffield, 2024.

4.6 The limitations of conjugate pairs

The main advantage of conjugate priors is that, when we can use them, Bayesian updates are simple to perform. Their main disadvantage is that, in many cases, we cannot use them. This can occur in two main ways:

- Our chosen model family does not have *any* conjugate priors.
- Our chosen model family does have conjugate priors, but there are no choices of prior parameters that result in a conjugate prior that matches our prior beliefs.

  This case is particularly likely to happen if we use a prior based on expert opinions or on earlier experimental work.

For that reason the modern approach to Bayesian learning largely relies on the computational techniques introduced in Chapter 8.

Exercise **6.6** shows that mixtures of conjugate priors can be handled with similar ease to conjugate priors. This can help if we need to manufacture a prior distribution to reflect particular properties, but it only provides enough help in a small fraction of situations.

©Nic Freeman, University of Sheffield, 2024.

4.7 Exercises on Chapter 4

You can find formulae for named distributions in Appendix A.

⋆ **4.1** (a) Consider the Bayesian model (X, Θ) with model family $M_\theta \sim \mathrm{N}(\theta, 2^2)^{\otimes 3}$ and prior $\Theta \sim \mathrm{N}(0, 1)$.

(i) Use Lemma 4.2.2 to find the posterior distribution given the data $x = (3.88,\ 2.34,\ 7.86)$, which satisfies $\sum_1^3 x_i = 14.08$.

(ii) Write down the probability density functions of the sampling and predictive distributions given by this model for a single data point. Give your answers in the form $\int_{\mathbb{R}} f_{\mathrm{N}(\cdot,\cdot)}(\cdot) f_{\mathrm{N}(\cdot,\cdot)}(\cdot)\, d(\cdot)$.

(b) Inside the files `2_dist_sketching.ipynb` and `2_dist_sketching.Rmd`, below the part corresponding to Exercise **3.1**, you will find code for sketching the sampling p.d.f. of the Bayesian model (X, Θ) from Example 3.2.3, with model family $M_\theta \sim \mathrm{Exp}(\lambda)$ and prior $\Lambda \sim \Gamma(2, 60)$. Modify the code given to sketch the sampling and predictive distributions from (a), on the same graph.

⋆⋆ **4.2** Let $\alpha, \beta > 0$. Let (X, Θ) be a discrete Bayesian model with model family $M_\theta \sim \mathrm{Geometric}(\theta)^{\otimes n}$ and parameter $\theta \in [0, 1]$. Suppose that the prior is $\Theta \sim \mathrm{Beta}(\alpha, \beta)$ and let $x = (x_1, \dots, x_n)$ where $x_i \in \{0, 1, \dots, \}$. Show that the posterior is

$$\Theta|_{\{X=x\}} \sim \mathrm{Beta}\left(\alpha + n,\ \beta + \sum_{i=1}^n x_i\right).$$

⋆⋆ **4.3** Let $\alpha, \beta > 0$. Let (X, Θ) be a discrete Bayesian model with model family $M_\theta \sim \mathrm{Poisson}(\theta)^{\otimes n}$ and parameter $\theta \in (0, \infty)$. Suppose that the prior is $\Theta \sim \mathrm{Gamma}(\alpha, \beta)$ and let $x = (x_1, \dots, x_n)$ where $x_i \in \{0, 1, \dots, \}$. Show that the posterior is

$$\Theta|_{\{X=x\}} \sim \mathrm{Gamma}\left(\alpha + \sum_{i=1}^n x_i,\ \beta + n\right).$$

⋆⋆ **4.4** Let $\mu \in \mathbb{R}$ and $\alpha, \beta > 0$. Let (X, T) be a discrete Bayesian model with model family $M_\theta \sim \mathrm{N}(\mu, \frac{1}{\tau})^{\otimes n}$ and parameter $\tau \in (0, \infty)$. Suppose that the prior is $T \sim \mathrm{Gamma}(\alpha, \beta)$ and let $x = (x_1, \dots, x_n)$ where $x_i \in (0, \infty)$. Show that the posterior is

$$T|_{\{X=x\}} \sim \mathrm{Gamma}\left(\alpha + \frac{n}{2},\ \beta + \frac{1}{2}\sum_{i=1}^n (x_i - \mu)^2\right).$$

⋆ **4.5** Using the model from Exercise **4.4**, with $\mu = 0$ and prior $\mathrm{Gamma}(2, 2)$, use a computer package of your choice to produce a graph of the prior and posterior density functions, given the data

$$x = (0.22,\ -0.17,\ 1.22,\ -0.13,\ 0.05,\ 0.79,\ -0.45,\ -0.30,\ 0.09,\ -0.16).$$

This data satisfies $\sum_1^{10} x_i^2 = 2.53$. Then, draw a second graph of the sampling and predictive density functions for a single data point.

©Nic Freeman, University of Sheffield, 2024.

4.6 Let $u \in \mathbb{R}$. Let (X, Θ) be a continuous Bayesian model with model family $M_\theta \sim \mathrm{N}(\theta, 2^2)^{\otimes 10}$ and parameter $\theta \in \mathbb{R}$. Suppose that the prior is $\Theta \sim \mathrm{N}(0, 2^2)$, and that we have data

$$x = (5.29,\ 1.20,\ 2.94,\ 6.72,\ 5.60,\ -2.93,\ 2.85,\ -0.45,\ -0.31,\ 1.23).$$

This data satisfies $\sum_1^{10} x_i = 133.19$.

⋆ (a) Use Lemma 4.2.2 to find the posterior distribution $\Theta|_{\{X=x\}}$.

⋆ (b) Using a computer package of your choice, implement the Bayesian update given by Lemma 4.2.2 with $n = 1$, and use it to perform 10 Bayesian update steps, one for each x_i, on the model (X', Θ) with model family $M_\theta \sim \mathrm{N}(\theta, 2)$. Write down the resulting posterior distribution.

⋆⋆ (c) What do you notice? Investigate this as you vary the data and prior parameters. Would the same thing happen with the other families of conjugate pairs in this chapter?

⋆⋆ **4.7** Let $a, b, \beta > 0$. Let (X, Θ) be a discrete Bayesian model with model family $M_\theta \sim \mathrm{Weibull}(\theta, \beta)^{\otimes n}$ and parameter $\theta \in (0, \infty)$. Suppose that the prior is $\Theta \sim \mathrm{IGamma}(a, b)$ and let $x = (x_1, \dots, x_n)$ where $x_i \in \{0, 1, \dots, \}$. Show that the posterior is

$$\Theta|_{\{X=x\}} \sim \mathrm{IGamma}\left(a + n,\ b + \sum_{i=1}^{n} x_i^\beta\right).$$

⋆ **4.8** Match each of Lemmas 4.1.5, 4.1.6, 4.2.1, 4.5.2 and Exercises **4.2**, **4.3**, **4.4**, **4.7** to their corresponding rows on the reference sheet of conjugate pairs given in Appendix A

⋆ **4.9** Recall the relation $\propto$ from Definition 4.1.1. Let f, g, h be functions with the same domain. Explain briefly why (a) $f \propto f$; (b) if $f \propto g$ then $g \propto f$; (c) if $f \propto g$ and $g \propto h$ then $f \propto h$.

(⊘) *Remark: These three properties are the definition of an equivalence relation.*

©Nic Freeman, University of Sheffield, 2024.

Chapter 5

The prior

We've spent most of our energy in Chapters 2-4 on understanding Bayesian models and performing the Bayesian update step. In Chapter 4 we focused on techniques for choosing the prior in a way that would make calculations straightforward to perform. In this chapter we maintain our focus on the prior, but with the opposite goal. Our interest here is in choosing a prior that best reflects a set of beliefs.

There are two parts to this chapter. Section 5.1 focuses on techniques for choosing a prior based on the opinions on experts. We are interested to quantify these opinions, in order to combine them with data, and we hope that by doing so we will produce more accurate results than could be obtained from the data alone. Sections 5.2 and 5.3 are concerned with the opposite situation where we want to focus solely on the data, and carry out our analysis with as few preconceptions as is possible.

Neither of these situations gives conjugate priors, in general. Consequently they lead to Bayesian updates that require computational methods, which we will study in Chapter 8. In both cases we must continue to abide by Cromwell's rule from Section 4.4: the chosen prior should allow a non-zero probability for all parameter values that are physically possible.

©Nic Freeman, University of Sheffield, 2024.

5.1 Elicitation

Elicitation is the process of extracting an individuals beliefs about some unknown quantity, and representing those beliefs via probabilities. It is a difficult and inexact process. People often struggle to turn their thoughts into probabilities and are susceptible to many different psychological biases. There is no reason to expect that a single persons beliefs will be self-consistent. We will discuss psychological biases in Section 5.1.1. For now let us focus on the process of elicitation.

The first question we need to answer is *whose* prior we actually want. For example, if we are trying to determine the effectiveness of a drug, should we use the prior beliefs of the pharmaceutical company, or perhaps the regulators, perhaps even the patients? As a general rule, the prior should represent the beliefs held by the person(s) who decides what actions should be taken in response to the statistical analysis. They should, ideally, be the same person(s) as will face the consequences of a poor or incorrect decision. They are known as the *elicitee* for the duration of the process.

Eliciting probabilities

We have a limited capacity to think in terms of probabilities. For example, no elicitee can use their personal experience to judge the difference between probability 0.5 and 0.5001. When people are prepared to state the probability of some event exactly, it is usually based on some sort of symmetry. For example most people will tell you that for a fair six sided dice we have $\mathbb{P}[\text{throw a } 6] = \frac{1}{6}$. In reality the probability is not exactly $\frac{1}{6}$, because the dice is not perfectly symmetric, but it is close enough for most practical purposes.

Except for symmetrical cases, elicitees estimating probabilities will generally rely on a mixture of their intuition and memory. We can help to understand their beliefs by choosing our questions carefully. It is good practice to focus on quantities that are meaningful to the elicitee, which usually means asking about quantities they have actually observed, or about relationships between quantities that they have expert knowledge of. In a complex model we may have to avoid asking directly about the parameters, because the elicitee may not understand what these parameters represent, even though our goal is to choose a prior distribution.

Eliciting distributions

We now consider how to elicit a whole distribution. Whole distributions are complicated objects and we can never claim that a certain distribution will perfectly represent someone's beliefs about an unknown quantity. The best we can hope for a is a reasonable representation.

We generally concentrate on two aspects of the distribution:

- **Location** represents the value, or range of values, where the unknown quantity is most likely to be. It might be a guess for a mode, median or mean.
- **Dispersal** represents the level of certainty that the parameter falls within its most likely range of values. It might be represented using a variance.

Non-statisticians will often have difficulty dealing directly with concepts like mode, median, mean and variance. Instead of asking directly, a common strategy is to choose a family of distributions and then elicit probabilities to determine appropriate parameters. We might start this process by asking the elicitiee to draw the rough shape of the distribution representing their beliefs, and choose a family able to reproduce that shape.

©Nic Freeman, University of Sheffield, 2024.

It is usually easier to elicit estimates of location than it is to elicit estimates of dispersal. An elicitee with a good understand of probability might be willing to specify percentiles directly, for example to state values of q such that $\mathbb{P}[\Theta \leq q] = 0.95$ and $\mathbb{P}[q \leq \Theta] = 0.95$, but many people will find this difficult, particularly for probabilities that are close to zero or one.

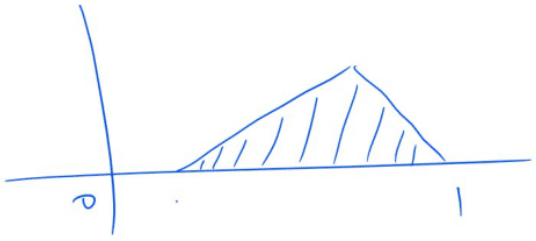
The following scheme is known as the *bisection method*. It focuses on events of equal probability that (according to the elicitee) have a good chance of occurring. It seeks to elicit information about a single unknown real parameter θ.

1. The elicitee is first asked to give a value m such that the events $\theta \in (-\infty, m]$ and $\theta \in [m, \infty)$ are equally likely.
2. Next, the elicitee is asked to give a value l such that the events $\theta \in (-\infty, l]$ and $\theta \in [l, m]$ are equally likely.
3. Lastly, the elicitee is asked to give a value u such that the events $\theta \in [m, u]$ and $\theta \in [u, \infty)$ are equally likely.

In more statistical language, the elicitee provides their estimate for the 25th percentile l, the median of 50th percentile m and the 75th percentile u. We could extend the process by splitting up further into more intervals of equal probability, or by using other percentiles, but we should be wary that increasing the complexity of the questions will also increase the risk that the elicitee fails to communicate their beliefs accurately.

We use the quantities l, m, u obtained to deduce the parameters, within our chosen family of possible prior distributions. It is helpful that we tend to have three quantities and (for named distributions) only one or two parameters, because this allows us to check up on how well we have represented the individuals prior beliefs.

Example 5.1.1 Suppose that we have carried out the bisection method and obtained $m = 0.7, l = 0.5, u = 0.8$, and the elicitee has sketched a distribution for a parameter $\theta \in [0, 1]$ that looks like this:

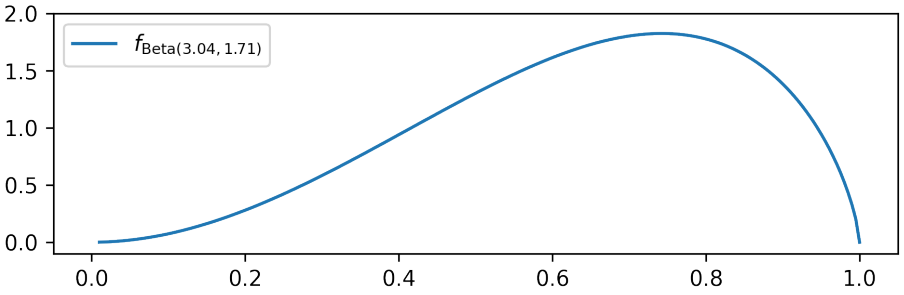


We decide to try and represent this information with a Beta(α, β) distribution. We solve the equations

$$\mathbb{P}[\text{Beta}(\alpha, \beta) \leq 0.3] = 0.25, \qquad \mathbb{P}[\text{Beta}(\alpha, \beta) \geq 0.8] = 0.25$$

numerically to obtain $\alpha = 3.04$ and $\beta = 1.71$. The median of the Beta$(3.04, 1.71)$ distribution is 0.66, also obtained numerically. This is close to the elicitees value for $m = 0.7$. The distribution we have obtained is:

©Nic Freeman, University of Sheffield, 2024.



It is a reasonable match for what the elicitee has drawn. It also accounts for Cromwell's rule by putting a small amount of probability on θ close to zero, which the elicitee did not think possible.

As part of the elicitation process we usually need to decide how strongly the prior should focus the values taken by θ into a particular region. A prior that strongly focuses θ on one (or occasionally more) small regions is known as a *strongly informative* prior. A prior that does not is generally known as a *weakly informative* prior. These are not mathematical definitions but they are very commonly used terms. You will often see them shortened to simply 'strong' and 'weak'.

©Nic Freeman, University of Sheffield, 2024.

5.1.1 Psychological biases

People often take decisions using heuristics, which are shortcuts that are used to make quick and effective guesses. For example people will often assume that a more expensive product will be of better quality, and may base their purchasing decisions partly on this idea; it is often, but not always, true. These sorts of heuristics can introduce biases into the elicitation process, when heuristics struggle to capture the reality of a complex situation. Some pitfalls to be aware of during elicitation:

1. **Availability bias.** This is where an elicitee overestimates the probability of an event because it is easy to remember (or notice) that event, or because it has recently occurred.

   For example, after hearing news of a plane crash, people are more susceptible to overestimate the frequency of plane crashes.

2. **Anchoring.** This is where an elicitee relies too heavily on a single piece of information.

3. **Hindsight bias.** This is when an elicitee falsely believes that they would have predicted an event, *after* that event has happened. This behaviour risks underestimating the probabilities of other outcomes.

4. **Overconfidence.** This is the tendency to give too much probability to events that are believed to be likely, and consequently underestimate the probability of unlikely events.

   For example, in many studies where people were asked to provide 95% intervals for various unknown quantities, the true value lay outside of the estimated intervals as much as 20-30% of the time.

There are many other ways in which the heuristics we rely on in everyday life can lead to errors and biases. It is never possible to eradicate them all, but it is clear that experience and training in making probabilistic statements, as well as making elicitees aware of potential sources of bias, tends to result in more accurate estimation.

©Nic Freeman, University of Sheffield, 2024.

5.2 Uninformative priors

If we are able to construct a prior distribution based on the opinions of experts, or on earlier research, then it will often be helpful to do so. It is not always possible, particularly if we are dealing with a situation in which very little is known, or if (for whatever reason) we wish to test how well expert opinions agree with the available data. In this section we discuss how to choose a prior that contains little presumption about what the best parameter values are. The general term for such priors is *uninformative*. Let us detail some approaches based on this idea.

Approach 1: uniform priors. Often, the best trick available here is the most obvious one. If the parameter space Π is a finite interval (for each parameter) then we can simply choose the uniform distribution. This makes our random choice Θ of the parameter equally likely to be anywhere within the parameter space Π. It is sometimes known as the 'principle of indifference', a term introduced by the economist Maynard Keynes in 1921.

Approach 2: improper priors. If the parameter space is an infinite interval, we cannot choose the uniform distribution, because there is no uniform distribution on an infinite interval! See Exercise 5.6. We need a new definition to understand this situation.

Definition 5.2.1 Let $\Pi \subseteq \mathbb{R}^d$ be the parameter space of a Bayesian model. A function $f : \Pi \to [0, \infty)$ such that $\int_\Pi f(x)\,dx = \infty$ is known as an *improper prior*, or more strictly an improper prior density function.

We use the term *proper* prior density function for the probability density functions of random variables that we could use for the prior. Note that if $\int_\Pi f(x)\,dx < \infty$ then we can define $\widetilde{f}(x) = \frac{1}{\int_\Pi f(y)\,dy} f(x)$ and then $\widetilde{f}$ is a prior density function with $f \propto \widetilde{f}$. Definition 5.2.1 captures the situation that we cannot turn f into a proper prior by including a normalizing constant. We use the same proper vs. improper terminology for posterior density functions, and for density functions in general.

For example, the functions

$$g(\theta) = \begin{cases} 1 & \text{for } \theta \in [0, \infty), \\ 0 & \text{otherwise,} \end{cases} \qquad \text{and } h(\theta) = \begin{cases} \frac{1}{\theta} & \text{for } \theta \in (0, 1], \\ 0 & \text{otherwise,} \end{cases}$$

are both improper density functions. You can check that $\int_\mathbb{R} g(\theta)\,d\theta = \int_\mathbb{R} h(\theta)\,d\theta = \infty$.

When Π is an infinite interval, a common approach is to use an improper prior $f_\Theta(\theta)$ and use Bayes rule anyway, in which case we obtain

$$f_{\Theta|_{\{X=x\}}}(\theta) \propto L_{M_\theta}(x) f_\Theta(\theta)$$

as usual. There are good theoretical reasons for doing so but they are beyond what we can cover in this course. We will still refer to Theorems 2.4.1 and 3.1.2 for these cases.

If $\int_\Pi L_{M_\theta}(x) f_\Theta(\theta)\,d\theta$ is finite then we can normalize to obtain $\widetilde{f}_{\Theta|_{\{X=x\}}}(\theta)$, which is still a p.d.f. corresponding to some random variable, and we can use that as our posterior. That situation does happen, but it is also possible that $\int_\Pi L_{M_\theta}(x)\,d\theta = \infty$. In this case we could take our improper posterior and use it as another improper prior in a future Bayesian update, and so on – but if we do not eventually reach a proper posterior density function (after some number of Bayesian update steps) then we will find it difficult to interpret the results of our analysis.

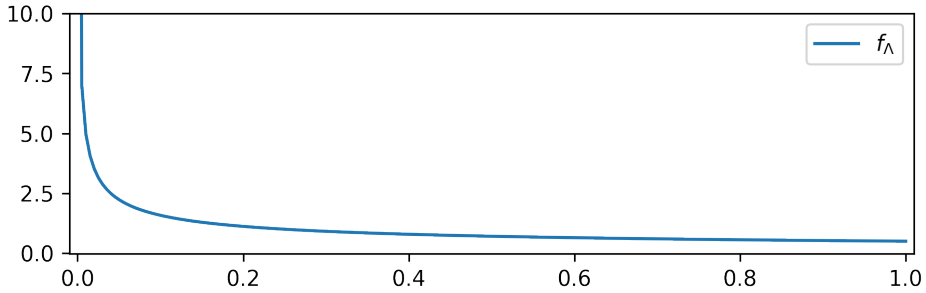
©Nic Freeman, University of Sheffield, 2024.

Approach 3: use a weak prior. To avoid getting involved with improper priors, a common technique is to choose a weak (but proper) prior distribution that is very well spread out across all plausible regions of the parameter space. We did this in Example 4.2.4, where we took $\Theta \sim N(30, 5^2)$ a normal distribution with a variance so large that it contained very little information about where the value of $\Theta \approx 30$ would be. A prior with infinite variance, such as the Cauchy distribution, is also an effective way to implement this idea. Note, however, that in many cases choosing a prior with this property will take us outside of the conjugate pairs from Chapter 4, and when that happens we will have to use numerical techniques (to be introduced in Chapter 8) to perform the Bayesian updates.

©Nic Freeman, University of Sheffield, 2024.

5.2.1 The philosophy of not knowing anything

Various philosophical arguments have been explored to try and make sense of the idea that a particular choice of prior corresponds to 'knowing nothing'. For example, consider the following argument:

> *A uniform prior* $\Theta \sim \text{Uniform}(0,1)$ *supposedly contains no preference for any value of* $\theta \in (0,1)$*, but if we use a different parametrization of our model family, say* $\lambda = \theta^2$ *then our prior becomes* $\Lambda = \Theta^2$ *with p.d.f.*
>
> $$f_\Lambda(\lambda) = \begin{cases} \frac{1}{2\sqrt{\lambda}} & \text{for } \lambda \in (0,1) \\ 0 & \text{otherwise,} \end{cases}$$
>
> 
>
>
> *and this is biased towards smaller values in* $[0,1]$*. If a prior really corresponds to having no preference for any value, then re-parametrization should not change that fact.*

This argument is non-mathematical and exploits the fact that its reader has no clear understanding of what 'no preference' means. If we consider more aggressive re-parametrizations too, say $\lambda = \theta^N$ for large N, then it doesn't really matter what continuous prior we start with, if $\Lambda \in [0,1]$ then samples of $\Lambda = \Theta^N$ will mostly be very close to zero i.e. nearly deterministic.

What we should take away from this is that Approach 1 and Approach 3 (from Section 5.2) are actually very similar. A slight re-parametrization of our model family will turn one approach into the other, and if we have no strong feeling about which prior is best then we won't have a strong feeling for which parametrization is best either. This is fine – re-parametrizing our model family slightly won't change the outcomes of our analysis much either.

Remark 5.2.2 (⊘) To give a mathematical treatment of the argument above, the concept of *entropy* becomes important. Loosely, entropy measures how different one distribution is to another, but it is only well-defined as a relative concept – of one distribution to another[1]. The argument above fails to account for the fact that all entropy is a relative entropy; we can make sense of the difference between two sets of preferences, and some preferences are stronger than others, but the concept of 'no preference' does not really exist.

[1] The terminology here is clouded by the fact that the term entropy is widely used as a shorthand for the relative entropy to the uniform distribution on an interval of $\mathbb{R}$ (or uniform measure, more generally).

©Nic Freeman, University of Sheffield, 2024.

5.3 Reference priors

An interesting response to the argument in Section 5.2.1 was given by the statistician Harold Jeffreys in 1946. It leads to a particular suggestion for the choice of prior. Suppose that two different people, Alice and Bob, construct a Bayesian model, with one parameter. Alice uses the model family $(M_\theta)_{\theta\in\Pi}$ and Bob use the model family $(M_\varphi)_{\varphi\in\Pi}$, where θ and φ are related by some function $h(\theta)=\varphi$, where $h:\Pi\to\Pi$ with $\Pi\subseteq\mathbb{R}$ and h is strictly monotone increasing and differentiable. That is, they use the 'same' model family, but parametrize it differently.

Alice will use prior p.d.f. f_1 and Bob will use prior p.d.f. f_2. This means that Alice constructs a model with sampling distribution

$$f_{X_1}(x)=\int_\Pi f_{M_\theta}(x)f_1(\theta)\,d\theta$$

and Bob constructs a model with sampling distribution

$$f_{X_2}(x)=\int_\Pi f_{M_{h(\theta)}}(x)f_2(\theta)\,d\theta.$$

Alice and Bob have never met each other, and in fact they do not even know that each other exists. Neither of them knows the function h.

This is where we come in. We write the statistics textbook, that both Alice and Bob will both read. They will choose their prior based on our instructions – the same instructions, for both people. Can we provide Alice and Bob with a way to choose their individual priors that will make their models equal i.e. so that $f_{X_1}(x)=f_{X_2}(x)$?

Remark 5.3.1 If Alice and Bob did meet each other, then Alice could tell Bob what her prior Θ was and by comparing notes they could work out the function h. Bob could then choose his prior to be the p.d.f. of $h(\Theta)$, where Θ is Alice's prior. This choice makes $f_{X_1}(x)=f_{X_2}(x)$, see Exercise 5.7.

Returning to the situation where Alice and Bob do not meet, the surprising answer to the problem is: yes, this is possible. The solution is that we should tell them *both* to use the prior

$$f(\theta)\propto\mathbb{E}\left[\left(\frac{d}{d\lambda}\log(L_{M_\lambda}(X))\right)^2\right]^{1/2}\qquad\text{where }(M_\lambda)\text{ is their chosen model family and }X\sim M_\lambda. \tag{5.1}$$

Let us not worry about how Jeffreys found this solution, and let us just show that it really works.

Proof that the solution works: (⊘) Alice writes down her prior $f_1(\theta)\propto\mathbb{E}[(\frac{d}{d\theta}\log(L_{M_\theta}(X)))^2]^{1/2}$ and Bob writes down his prior, $f_2(\varphi)\propto\mathbb{E}[(\frac{d}{d\varphi}\log(L_{M_{h(\varphi)}}(X)))^2]^{1/2}$. Then Alice's model is

$$f_{X_2}(x)\propto\int_\Pi f_{M_\theta}(x)f_1(\theta)\,d\theta.$$

Alice doesn't know the function h, but substituting $\theta=h(\lambda)$, her model is equal to

$$\begin{aligned}f_{X_1}(x)&\propto\int_\Pi f_{M_{h(\lambda)}}(x)f_1(h(\lambda))h'(\lambda)\,d\lambda\\&\propto\int_\Pi f_{M_{h(\lambda)}}(x)f_1(h(\lambda))h'(\lambda)\,d\lambda\end{aligned}$$

©Nic Freeman, University of Sheffield, 2024.

$$\propto \int_\Pi f_{M_{h(\theta)}}(x) f_1(h(\theta)) h'(\theta)\, d\theta.$$

In the last line we have simply changed notation by setting $\lambda = \theta$. Meanwhile, Bob's model is equal to

$$f_{X_1}(x) \propto \int_\Pi f_{M_{h(\theta)}}(x) f_2(\theta)\, d\theta.$$

It follows that $f_{X_1}(x) = f_{X_2}(x)$ if we have $f_2(\theta) \propto f_1(h(\theta))h'(\theta)$. We will now show that this equation holds, for any strictly monotone function h. Writing $\varphi = h(\theta)$,

$$\begin{aligned} f_2(\varphi)^2 &\propto \mathbb{E}\left[\left(\frac{d}{d\varphi}\log(L_{M_\varphi}(X))\right)^2\right] \\ &\propto \mathbb{E}\left[\left(\frac{d}{d\theta}\log(L_{M_{h(\theta)}}(X)) \times \frac{d\varphi}{d\theta}\right)^2\right] \\ &\propto \mathbb{E}\left[\left(\frac{d}{d\theta}\log(L_{M_{h(\theta)}}(X)) \times h'(\theta)\right)^2\right] \\ &= f_1(h(\theta))^2 h'(\theta)^2. \end{aligned}$$

To reach the second line we use the chain rule. Taking square roots, we obtain that $f_{X_1}(x) = f_{X_2}(x)$ as required. ■

In the earlier half of the 20th the argument in Section 5.2.1 was taken quite seriously, and treated as a major philosophical reason to question the reliability of Bayesian statistics. In particular, the objection was that if both Alice and Bob tried to use the same uninformative prior but used models that were parametrized differently then they would obtain different results, despite having the same intentions and, from their own perspectives, the same methodology. Jeffreys showed that this difficulty could be entirely avoided with a particular choice of prior.

These arguments took place before modern computers, when it was difficult to test how well Bayesian methods worked in practice (except for conjugate priors). We are now better able to test how much different modelling errors matter. Statisticians today no longer attach much weight to this objection.

Starting from the ideas above and those in Remark 5.2.2, there is a modern branch of statistics that investigates uninformative priors with particular theoretical properties. A modern approach is to use a prior that tries to maximise the difference (in some sense) between the prior and posterior distribution, essentially seeking to maximise the influence of the data. Priors with this property are known as reference priors. Their theory is beyond what we can cover here, but it turns out that if we have only one parameter and we model our data as i.i.d. samples then the reference prior is the same as the prior proposed by Jeffreys – in more complicated cases, they are different. Let us investigate what the reference prior looks like for some particular choices of one-parameter model.

Definition 5.3.2 Suppose that (M_θ) is a family of distributions with parameter space $\Pi \subseteq \mathbb{R}$. The *reference prior* Θ associated (M_θ) has p.d.f. given by

$$f_\Theta(\theta) \propto \mathbb{E}\left[\left(\frac{d}{d\theta}\log(L_{M_\theta}(X))\right)^2\right]^{1/2}. \tag{5.2}$$

©Nic Freeman, University of Sheffield, 2024.

There are caveats to this definition. The reference prior might be an improper prior, or if the expectation in (5.2) is not finite then the reference prior may not exist.

Sometimes the reference prior Θ for (M_θ) is easier to find via the equation

$$f_\Theta(\theta) \propto \mathbb{E}\left[-\frac{d^2}{d\theta^2}\log(L_{M_\theta}(X))\right]^{1/2} \tag{5.3}$$

which is equivalent to (5.2).

Remark 5.3.3 (⊘) To deduce (5.3) from (5.2), use the partial differentiation identity $\frac{\partial^2}{\partial\theta^2}\log f(x,\theta) = \frac{1}{f(x,\theta)}\frac{\partial^2}{\partial\theta^2}f(x,\theta) - \left(\frac{\partial}{\partial\theta}\log f(x,\theta)\right)^2$ and that $\mathbb{E}\left[\frac{1}{f(X;\theta)}\frac{\partial^2}{\partial\theta^2}f(X,\theta)\,\middle|\,\theta\right] = \frac{\partial^2}{\partial\theta^2}\int_{\mathbb{R}} f(x,\theta)\,dx = 0$. We omit the details.

Example 5.3.4 For the Bernoulli model family $(M_p)_{p\in[0,1]}$ where $M_p \sim \text{Bernoulli}(p)$, the likelihood is

$$L_{M_p}(x) = \begin{cases} p^x(1-p)^{1-x} & \text{for } p\in[0,1] \\ 0 & \text{otherwise} \end{cases}$$

for $x \in \{0,1\}$. Hence $\frac{d}{dp}\log(L_{M_p}(x)) = \frac{p}{dp}(x\log p + (1-x)\log(1-p)) = \frac{x}{p} + \frac{1-x}{1-p}$. For $p\in[0,1]$, the density function of the reference prior P is given by

$$\begin{aligned} f_P(p) &\propto \mathbb{E}\left[\left(\frac{d}{dp}\log(L_{M_p}(X))\right)^2\right]^{1/2} \\ &\propto \mathbb{E}\left[\left(\frac{X}{p}+\frac{1-X}{1-p}\right)^2\right]^{1/2} \\ &\propto \left(p\left(\frac{1}{p}+\frac{1-1}{1-p}\right)^2 + (1-p)\left(\frac{0}{p}+\frac{1-0}{1-p}\right)^2\right)^{1/2} \\ &\propto \left(\frac{1}{p}+\frac{1}{1-p}\right)^{1/2} \\ &\propto p^{-1/2}(1-p)^{-1/2}. \end{aligned}$$

Using Lemma 1.2.5 we recognize that $P \sim \text{Beta}(\frac{1}{2},\frac{1}{2})$.

A useful fact is that the reference prior for M_θ and $M_\theta^{\otimes n}$ are identical, in the sense that they are $\propto$ to each other. This is shown in Exercise **5.8**.

©Nic Freeman, University of Sheffield, 2024.

5.4 Exercises on Chapter 5

⋆ **5.1** (a) Let X be the age in years of a person sampled uniformly at random from the UK population. Write down your best guess at $\mathbb{P}[X \geq x]$ for the values

$$x = 10, 20, 30, 40, 50, 60, 70, 80, 90.$$

(b) Sketch the distribution that you obtained in (a) as a histogram, of $\mathbb{P}[X = x]$ for the values of x above. Does the histogram accurately represent your prior beliefs about the UK population? If not, make changes until you think it does.

(c) In the solutions to this question you will find a table of these statistics, obtained in the UK Census 2021. For each value of x, calculate the value of $\frac{\text{your estimate}}{\text{census value}}$ (for example, if this value is 2, your estimate was twice the true value). For which values of x was your prior distribution least accurate?

⋆⋆ **5.2** Let τ denote the maximum temperature that will occur outdoors in Sheffield tomorrow.

(a) Sketch your prior density for τ.

(b) (i) Perform the bisection method on yourself (or do this question with a friend) to elicit your 50th, 25th and 75th percentiles for τ. Use the 25th and 75th percentiles to construct a Normal distribution $\mathrm{N}(\mu, \sigma^2)$ representing your beliefs. How close is μ to your 50th percentile?
To help with this, the code used in Example 5.1.1 can be found within `5_elicitation_example.ipynb` and `5_elicitation_example.Rmd`.

(ii) Without using your answer to (i), write down your estimation of the 5th and 95th percentiles for τ.

(iii) Compare your answers to (ii) to the implied probabilities of the distribution you found in (i). Is the Gaussian distribution a good fit for you beliefs?

(c) Repeat part (b) using the Cauchy distribution instead of the Normal distribution. Which family of distributions better fits your beliefs?

⋆⋆ **5.3** Let $(M_\lambda)_{\lambda \in (0,\infty)}$ be the Poisson model family, in which $M_\lambda \sim \text{Poisson}(\lambda)$. Show that the reference prior of this model family is given by

$$f_\Lambda(\lambda) \propto \begin{cases} \lambda^{-1/2} & \text{for } \lambda > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Does this define a proper prior or an improper prior?

⋆⋆ **5.4** We will model the monthly occurrence of fires within a nameless small town using the Poisson(λ) model family. As very little is known prior to the data collection, we will use the reference prior Λ obtained in Exercise **5.3**.

The number of fires recorded during the past 12 months is $x = (x_1, \ldots, x_{12}) \in \{0, 1, \ldots, \}^{12}$.

(a) Suppose that $x = (0, 1, 0, 0, 2, 0, 0, 0, 1, 0, 1, 0)$. Find and sketch the distribution of the posterior $\Lambda|_{\{X=x\}}$.

(b) Show that $\Lambda|_{\{X=x\}}$ is a proper distribution for all possible values of x.

©Nic Freeman, University of Sheffield, 2024.

⋆⋆ **5.5** Let (X, Θ) be a Bayesian model with model family $M_\theta \sim \text{Uniform}(0, \theta)$, the continuous uniform distribution on $(0, \theta)$.

(a) Given the prior $\Theta \sim \text{Exp}(1)$, find the posterior density function $f_{\Theta|_{\{X=x\}}}(\theta)$.
You should discover that $f_{\Theta|_{\{X=x\}}}(\theta) = 0$ for $\theta < x$. Can you explain (without reference to your calculations) why this has happened?

(b) Instead, let us now use the prior $\Theta \sim \text{Uniform}(1, 2)$, and suppose that our data is $x = 3$. Is the posterior distribution well defined? What has gone wrong here?

⋆⋆⋆ **5.6** We say that a random variable U is uniformly distributed on an interval I if, for all $a < b$ and $c > 0$ such that both $[a, b] \subseteq I$ and $[a+c, b+c] \subseteq I$, we have $\mathbb{P}[U \in [a, b]] = \mathbb{P}[U \in [a+c, b+c]]$.
Show that there is no random variable U that is uniformly distributed on $[0, \infty)$.

⋆⋆⋆ **5.7** Prove the claim in Remark 5.3.1. You should start by finding an expression for the p.d.f. of $h(\Theta)$, where Θ has p.d.f. f_1.

⋆⋆⋆ **5.8** Let (M_θ) be a family of distributions and let $n \in \mathbb{N}$. Let $f_{M_\theta}(\theta)$ denote its reference prior and let $f_{M_\theta^{\otimes n}}(\theta)$ denote the reference prior of the family $(M_\theta^{\otimes n})$. Show that $f_{M_\theta}(\theta) \propto f_{M_\theta^{\otimes n}}(\theta)$.

©Nic Freeman, University of Sheffield, 2024.

Chapter 6

Discussion

We have now understood enough about Bayesian inference to discuss how it compares to other techniques. We will do so in Section 6.2. We first give an outline of the various different notations that are used for the Bayesian framework, most of which are more condensed than the notation we have used in Chapters 1-5.

©Nic Freeman, University of Sheffield, 2024.

6.1 Bayesian shorthand notation

Recall that for a random variable Y we define the likelihood function L_Y by

$$L_Y(y) = \begin{cases} p_Y(y) & \text{where } p_Y \text{ is the p.m.f. and } Y \text{ is discrete,} \\ f_Y(y) & \text{where } f_Y \text{ is the p.d.f. and } Y \text{ is continuous.} \end{cases} \tag{6.1}$$

We continue with our convention of denoting probability density functions by f and probability mass functions by p.

This notation allows us to write the key equation from Theorems 2.4.1 and 3.1.2, for the distribution of the posterior, in a single form. If (X, Θ) is a (discrete or continuous) Bayesian model, where Θ is a continuous random variable with p.d.f. $f_\Theta(\theta)$, and the model family (M_θ) has likelihood function M_θ, then the posterior distribution of Θ given the data x has p.d.f.

$$f_{\Theta|_{\{X=x\}}}(\theta) = \frac{L_{M_\theta}(x) f_\Theta(\theta)}{L_X(x)} \tag{6.2}$$

where $Z = L_X(x)$ is the normalizing constant, or equivalently $f_{\Theta|_{\{X=x\}}}(\theta) \propto L_{M_\theta}(x) f_\Theta(\theta)$. In both Sections 2.2 and 3.1 we noted that $M_\theta \overset{\mathrm{d}}{=} X|_{\{\Theta=\theta\}}$, which leads to

$$f_{\Theta|_{\{X=x\}}}(\theta) \propto L_{X|_{\{\Theta=\theta\}}}(x) f_\Theta(\theta). \tag{6.3}$$

The term $L_{X|_{\{\Theta=\theta\}}}(X)$ is often known as the likelihood function of the Bayesian model and equation 6.2 is yet another version of Bayes' rule. It is the most general version of the Bayes' rule that we will encounter within this course, and it is the basis for most practical applications of Bayesian inference.

Some textbooks, and many practitioners, prefer to use a more condensed notation for equations (6.2) and (6.3). They write simply $f(y)$ for the likelihood function of Y, and $f(x)$ for the likelihood function of X. Conditioning is written as $f(y|x)$ for the likelihood function of Y given the event $\{X = x\}$. This notation requires that we must only ever write x for samples of X and y for samples of Y, or else we would not be able to infer which random variables were involved. In this notation (3.1) becomes

$$f(\theta|x) = \frac{f(x|\theta) f(\theta)}{f(x)}, \tag{6.4}$$

which is easy to remember! It bears a close similarly to $\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\mathbb{P}[B]}{\mathbb{P}[A]}$, which is Bayes' rule for events. Note that in (6.4) the 'function' f is really representing four different functions, dependent upon which variable(s) are fed into it – that part of the notation can easily become awkward and/or confusing if you are not familiar with it.

There are many variations on the notation in (6.4).

1. Some textbooks prefer to denote likelihood by $p(\cdot)$ instead of $f(\cdot)$, giving $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$. Some use a different symbol for likelihood functions connected to Θ and those connected to X, for example $p(\theta|X) = \frac{l(x|\theta)p(\theta)}{l(x)}$ where p denotes prior and posterior and l denotes the likelihood of the model family.

2. Sometimes the likelihood is omitted entirely, by writing e.g. $\theta|x$ to denote the distribution of the posterior $\Theta|_{\{X=x\}}$. Here lower case letters are used to blur the distinction between random variables and data. For example, you might see a Bayesian model with Binomial model family and a Beta prior defined by writing simply $\theta \sim \mathrm{Beta}(a, b)$ and $x|\theta \sim \mathrm{Bin}(n, \theta)$.

©Nic Freeman, University of Sheffield, 2024.

3. Some textbooks use subscripts to indicate which random variables are conditioned on, as we have done, but in a slightly different way e.g. $f_{X|Y}(x, y)$ instead of $f_{X|_{\{Y=y\}}}(x)$.

In this course we refer to all these various notations as *Bayesian shorthand*, or simply shorthand.

Using shorthand can make Bayesian statistics very confusing to learn, so we have avoided it so far within this course. We will sometimes use it from now on, when it is convenient and clear in meaning. This includes several of the exercises at the end of this chapter. For those of you taking MAS61006, Bayesian shorthand will be used extensively in the second semester of your course. Hopefully, by that point you will be familiar enough with the underlying theory that they will save you time rather than cause confusion.

Remark 6.1.1 You should write your answers to questions in the same style of notation as the question uses, unless you are explicitly asked to do otherwise.

©Nic Freeman, University of Sheffield, 2024.

6.1.1 A technical remark

Remark 6.1.2 (⊘) The underlying reason for most of the troubles with notation is that, from a purely mathematical point of view, there is no need to restrict to the two special cases of discrete and continuous distributions. It is more natural to think of both Theorems 2.4.1 and 3.1.2 as statements of the form 'we start with a distribution (the prior) and we perform an operation to turn it into another distribution (the posterior)'. The operation involved here (the Bayesian update) can be made sense of in a consistent way for all distributions, but it requires the *disintegration theorem* which takes some time to understand.

At present, the strategy chosen by most statisticians is to simply not study disintegration. This is partly for historical reasons. It was clear how to do the continuous case several decades before the disintegration theorem was proved, and the discrete case was understood two centuries before that. Based on these two special cases statisticians developed the idea of a likelihood function, split into two cases as in (6.1). The use of likelihood functions then became well established within statistics, before disintegrations of general distributions were understood by mathematicians. Consequently statistics generally restricts itself to the discrete and continuous cases that we have described in this course.

There are advantages and disadvantages to this choice. It still gives us enough flexibility to write down most of the Bayesian models that we might want to use in data analysis – although we would struggle to handle a model family that uses e.g. the random variable of mixed type in Exercise **1.2**. Very occasionally it makes things actually go wrong, as we noted in Remark 3.1.3. The main downside is that we often have to treat the discrete and continuous separately, as we did in Chapters 2 and 3. That consumes a bit of time and it leaves us with a weaker understanding of what is going on.

©Nic Freeman, University of Sheffield, 2024.

6.2 The connection to maximum likelihood

You have already seen maximum likelihood based methods for parameter inference in previous courses. They rely on the idea that, if we wish to estimate the parameter θ, we can use that value

$$\hat{\theta} = \arg\max_{\theta \in \Pi} L_{M_\theta}(x) \tag{6.5}$$

Here (M_θ) is a family of models and x is data, and we believe that for some value(s) of the parameter the model M_θ is reasonably similar to whatever physical process generated our data.

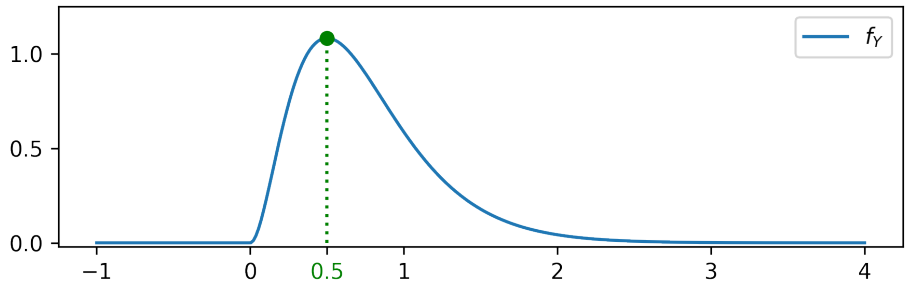
The value of $\hat{\theta}$, which is usually uniquely specified by (6.5), is known as the *maximum likelihood estimator* of θ, given the data x and model family (M_θ). Graphically, it is the value of θ corresponding to the highest point on the graph $\theta \mapsto L_{M_\theta}(x)$. Heuristically, it is the value of θ that produces a model M_θ that has the highest probability (within our chosen model family) to generate the data that we actually saw.

Recall that, for a discrete random variable Y, the *mode* is most likely single value for Y to take, or $\arg\max_{y \in R_Y} \mathbb{P}[Y = y]$ in symbols. You may be familiar with the following definition already, but we are about to need it, so we recall:

Definition 6.2.1 Let Y be a continuous random variable with range R_Y. The *mode* of Y is the value $y \in R_Y$ that maximises the p.d.f. $f_Y(y)$, given by $\arg\max_{y \in R_Y} f_Y(x)$.

Example 6.2.2 For continuous random variables, $\mathbb{P}[Y = y] = 0$ for all y. The idea here is that in this case the concept of 'most likely value' is best represented by the maximum of the probability density function. Let $Y \sim \text{Gamma}(3, 4)$, with p.d.f.

$$f_Y(y) = \begin{cases} 32y^2 e^{-3y} & \text{for } y > 0, \\ 0 & \text{otherwise.} \end{cases}$$



The mode is shown at its value $y = \frac{1}{2}$. This value can be found by solving the equation $\frac{df_Y(y)}{dy} = 32\left(2ye^{-4y} + y^2(-4e^{-4y})\right) = 32ye^{-4y}(2 - 4y) = 0$ and checking that the solution $y = \frac{1}{2}$ corresponds to a local maxima.

©Nic Freeman, University of Sheffield, 2024.

Comparing equations (6.5) and (6.3), there is a clear connection between MLEs and Bayesian inference: *if we take a flat prior (i.e. $f_\Theta(\theta)$ is constant) then the MLE $\hat{\theta}$ is equal to the mode of the posterior distribution.* This allows us to view the MLE approach as a simplification of the Bayesian approach. There are two steps to this simplification, to obtain the MLE approach from the Bayesian one:

1. We fix the prior to be a uniform distribution (or an improper flat prior, if necessary).
2. Instead of considering the posterior distribution as a random variable, we approximate the posterior distribution with a point estimate: its mode.

In principle we might make either one of these simplifications without the other one, but they are commonly made together. We are now able to discuss how the two approaches compare:

- We've seen in many examples that, as the amount of data that we have grows, the posterior distribution tends to become more and more concentrated around a single value. In such a case, the MLE becomes a very good approximation for the posterior. This situation is common when we have plenty of data – see Section 6.2.1 for a more rigorous (but off-syllabus) discussion.
- If we do not have lots of data then the approximation in step 2 will be less precise, and in the Bayesian case the influence of the prior will matter. In this situation whether Bayesian or MLE methods perform best depends on several factors.

  Bayesian methods require more work to implement, but they allow us to incorporate prior beliefs (if we have them). These prior beliefs can make the analysis more reliable, if they are realistic beliefs, but can make it less reliable if they are not. Bayesian methods generate a posterior distribution that has a clear meaning in terms of conditional probability, with little scope for misinterpretation.

  MLE based methods are comparatively easier to implement, but come with a risk of loss of detail from the approximation in step 2. They produce a point estimate for the known parameters, which is easier to communicate, but is also more open to misinterpretation. (We will discuss this issue in more detail in Chapter 7.)
- If our model is not a reasonable reflection of reality, or if having more data does not help us infer parameters more accurately, then *both* methods become unreliable – no matter how much data we have.

Remark 6.2.3 When we give a point estimate of a random variable it is more common to use the mean, but for the MLE we use the mode. The reason for doing so is simply that the mode often gives a nicer formulae.

Statistical methods based on MLEs and the simplifications 1 and 2 listed above are often known as 'frequentist' or 'classical' methods. You will sometimes find that statisticians describe themselves as 'Bayesian' or 'frequentist', carrying the implication that they prefer to use one family of methods over the other. This may come from greater experience with one set of methods or from a preference due to the specifics of a particular model.

To a great extent this distinction is historical. During the middle of the 20^{th} century methods based on simplifications 1 and 2 dominated statistics, because they could be implemented without the need for modern computers. Once it was realized that modern computers made Bayesian

©Nic Freeman, University of Sheffield, 2024.

methods possible (with complicated model families) the community that investigated these techniques needed a name and an identity, to distinguish itself as something new. The concept of identifying as 'Bayesian' or 'frequentist' is essentially a relic of that social process, rather than anything with a clear mathematical foundation.

Modern statistics makes use of both posterior distributions and (MLE or otherwise) simplifications of the posterior distribution. Sometimes it mixes the two approaches together, or chooses between them for model-specific reasons. We do need to divide things up in order to learn them, so will only study Bayesian models within this course – but in general you should maintain an understanding of other approaches too.

©Nic Freeman, University of Sheffield, 2024.

6.2.1 Making the connection precise (⊘)

Several theorems are known which actually prove, under wide ranging conditions, that when we have plenty of data the MLE and Bayesian approaches become essentially equivalent. These theorems are complicated to state, but let us give a brief explanation of what is known here.

Take a model family $(M_\theta)_{\theta\in\Pi}$ and define a Bayesian model (X, Θ) with model family $M_\theta^{\otimes n}$. This model family represents n i.i.d. samples from M_θ. Fix some value $\theta^* \in \Pi$, which we think of as the true value of the parameter θ. Let x be a sample from $M_{\theta^*}^{\otimes n}$. We write the posterior $\Theta|_{\{X=x\}}$ as usual.

Let $\hat{\theta}$ be the MLE associated to the model family $M_\theta^{\otimes n}$ given the data x, that is $\hat{\theta} = \arg\max_{\theta\in\Pi} L_{M_\theta^{\otimes n}}(x)$. Then as $n \to \infty$ it holds that

$$\Theta|_{\{X=x\}} \overset{\mathrm{d}}{\approx} \mathrm{N}\left(\theta^*, \frac{1}{n} I(\theta^*)^{-1}\right) \tag{6.6}$$

where $I(\theta)$ is the *Fisher information matrix* defined by $I(\theta)_{ij} = \mathbb{E}[\frac{\partial^2}{\partial\theta_i\theta_j} \log f_{M_\theta^{\otimes n}}(X)]$. The key point is that (6.6) says that the posterior $\Theta|_{\{X=x\}}$ and the MLE θ^* are in fact very similar, for large n, because of the factor $\frac{1}{n}$ in the variance.

Equation (6.6) is known as Laplace approximation. The mathematically precise form of this approximation, which replaces $\overset{\mathrm{d}}{\approx}$ in (6.6) by the concept of convergence is distribution, is known as the Bernstein von-Mises theorem. The first rigorous proof was given by Doob in 1949 for the special case of finite sample spaces. It has since been extended under more general assumptions, notably to cover countable state spaces, but the general case (and whatever conditions it may need) is still unknown. From these results we do know that various conditions are required for (6.6) to hold. We can also identify cases in which (6.6) will fail: for example if $M_\theta \sim \mathrm{Cauchy}(\theta, 1)$ then all of the terms in $I(\theta)$ will be undefined.

©Nic Freeman, University of Sheffield, 2024.

6.3 Exercises on Chapter 6

⋆ **6.1** Show that the mode of the Gamma(α, β) distribution is $\frac{\alpha-1}{\beta}$, where $\alpha \geq 1$. What about $\alpha \in (0, 1)$?

⋆⋆ **6.2** The following equations, written in Bayesian shorthand, are the key conclusions from results in earlier chapters of these notes. Which results are they from?

(a) $f(x|y) = \frac{f(y,x)}{f(y)}$.

(b) If $\theta \sim \text{Beta}(\alpha, \beta)$ and $x|\theta \sim \text{Bernoulli}(\theta)^{\otimes n}$ then $\theta|x \sim \text{Beta}(\alpha + k, \beta + n - k)$, where $x = (x_i)_1^n$ and $k = \sum_1^n x_i$.

Write the following results in Bayesian shorthand, using similar notation to that in parts (a) and (b).

(c) Lemma 4.2.1.

(d) From Section 4.5, the two facts above Lemma 4.5.2 concerning marginal and conditional distributions of the NGamma distribution.

⋆⋆ **6.3** The following results are written in Bayesian shorthand.

(a) If $x \sim N(0, 1)$ then $x|\{x > 0\} \sim |x|$.

(b) If x and y are independent then $x|y \sim x$.

In each case, write a version of the results in precise mathematical notation. Which parts of Chapter 1 are they closely related to?

⋆⋆ **6.4** Suppose that we model $x|\theta \sim \text{NegBin}(m, \theta)^{\otimes n}$, where $m \in \mathbb{N}$ is fixed and $\theta \in (0, 1)$ is an unknown parameter.

(a) Show that $f(x|\theta) \propto \theta^{mn}(1-\theta)^{\sum_1^n x_i}$.

(b) Show that the prior $\theta \sim \text{Beta}(\alpha, \beta)$ is conjugate to $\text{NegBin}(m, \theta)^{\otimes n}$, and find the posterior parameters.

(c) (i) Show that the reference prior for θ is given by $f(\theta) \propto \theta^{-1}(1-\theta)^{-1/2}$.

(ii) Does $f(\theta)$ define a proper distribution?

(iii) Find the posterior density $f(\theta|x)$ arising from this prior.

Hint: The setup given is a Bayesian model with model family $M_\theta \sim \text{NegBin}(m, \theta)^{\otimes n}$.

6.5 Suppose that we model $x|\mu, \tau \sim \text{N}(\mu, \frac{1}{\tau})^{\otimes n}$, where both μ and τ are unknown parameters. We use the improper prior $f(\mu, \tau) \propto \frac{1}{\tau}$ for $\tau > 0$, and $f(\tau) = 0$ elsewhere.

⋆⋆ (a) Show that for $\mu \in \mathbb{R}$ and $\tau > 0$ the posterior distribution satisfies

$$f(\mu, \tau|x) \propto \tau^{\frac{n}{2}-1} \exp\left(-\frac{\tau}{2}\sum_{i=1}^n (x_i - \mu)^2\right).$$

⋆⋆⋆ (b) Find the marginal p.d.f of $\tau|x$. Show that $(\mu, \tau)|x$ is a proper distribution if and only if $n \geq 2$.

©Nic Freeman, University of Sheffield, 2024.

Hint: The setup given is a Bayesian model with model family $M_{\mu,\tau} \sim \mathrm{N}(\mu, \frac{1}{\tau})^{\otimes n}$. *For part (b) use the sample-mean-variance identity* (4.10).

⋆⋆ **6.6** Let $(M_\theta)_{\theta\in\Pi}$ be a continuous family of distributions. For $i = 1, 2$, let Θ_i be a continuous random variable with p.d.f. f_{Θ_i}, both taking values in $\mathbb{R}^d$. Let $\alpha, \beta \in (0, 1)$ be such that $\alpha + \beta = 1$.

(a) Show that $f_\Theta(\theta) = \alpha f_{\Theta_1}(\theta) + \beta f_{\Theta_2}(\theta)$ is a probability density function.

(b) Consider Bayesian models (X_1, Θ_1) and (X_2, Θ_2), with the same model family (M_θ) and different prior distributions. Consider also a third Bayesian model (X, Θ) with model family (M_θ) and prior Θ with p.d.f. $f_\Theta(\theta) = \alpha f_{\Theta_1}(\theta) + \beta f_{\Theta_2}(\theta)$.

Show that the posterior distributions of these three models satisfy

$$f_{\Theta|_{\{X=x\}}}(\theta) = \alpha' f_{\Theta_1|_{\{X_1=x\}}}(\theta) + \beta' f_{\Theta_2|_{\{X_2=x\}}}(\theta)$$

where $\alpha' = \frac{\alpha Z_1}{\alpha Z_1 + \beta Z_2}$ and $\beta' = \frac{\beta Z_2}{\alpha Z_1 + \beta Z_2}$. Here Z_1 and Z_2 are the normalizing constants given in Theorem 3.1.2 for the posterior distributions of (X_1, Θ_1) and (X_2, Θ_2).

(c) Outline briefly how to modify your argument in (c) to also cover the case of discrete Bayesian models.

⋆⋆⋆ **6.7** *This question explores the idea in Exercise* ***4.6*** *further, but except for (a)(ii) it does not depend on having completed that exercise.*

(a) Let (M_θ) be a discrete or absolutely continuous family with range R. Let (X, Θ) be a Bayesian model with model family $M_\theta^{\otimes n}$. Let $x \in R^n$ and write $x(1) = (x_1, \dots, x_{n_1})$, $x(2) = (x_{n_1+1}, \dots, x_n)$. Let (X_1, Θ) and $(X_2, \Theta|_{\{X_1=x(1)\}})$ be Bayesian models with model families $M_\theta^{\otimes n_1}$ and $M_\theta^{\otimes n_2}$, where $n_1 + n_2 = n$.

(i) Show that

$$(\Theta_1|_{\{X_1=x(1)\}})|_{\{X_2=x(2)\}} \overset{\mathrm{d}}{=} \Theta|_{\{X=x\}}.$$

Use likelihood functions to write your argument in a way that covers both the discrete and absolutely continuous cases.

(ii) What is the connection between this fact and Exercise **4.6**?

(b) Rewrite your solution to (a)(i) in a Bayesian shorthand notation of your choice.

©Nic Freeman, University of Sheffield, 2024.

Chapter 7

Testing and parameter estimation

In this chapter we discuss aspects of statistical testing and parameter inference, using the Bayesian models set up in earlier chapters. Throughout this chapter we work in the situation of a discrete or absolutely continuous Bayesian model (X, Θ), where we have data x and posterior $\Theta|_{\{X=x\}}$. We keep all of our usual notation: the parameter space is Π, the model family is $(M_\theta)_{\theta\in\Pi}$, and the range of the model is R. Note that M_θ could have the form $M_\theta \sim (Y_\theta)^{\otimes n}$ for some random variable Y_θ with parameter θ, corresponding to n i.i.d. data points.

We have noted in Chapter 5 that an well chosen prior distribution can lead to a more accurate posterior distribution. Statistical testing is often used in situations where multiple different perspectives are involved and this makes the specification of prior beliefs more complicated. For example, trials of medical treatments involve patients, pharmaceutical companies and regulators, all of whom have different levels of trust in each other as wel as potentially different prior beliefs. It is common practice to check how much the results of statistical tests depend upon the choice of prior, often by varying the prior or comparing to a weakly informative prior.

©Nic Freeman, University of Sheffield, 2024.

7.1 Hypothesis testing

Hypothesis testing is surprisingly simple within the Bayesian framework. We first need to introduce the way to present the results.

Definition 7.1.1 Let A and B be events such that $\mathbb{P}[A \cup B] = 1$ and $A \cap B = \emptyset$. The *odds ratio* of A against B is

$$O_{A,B} = \frac{\mathbb{P}[A]}{\mathbb{P}[B]}.$$

It expresses how much more likely A is than B. For example, $O_{A,B} = 2$ means that A is twice as likely to occur than B; if $O_{A,B} = 1$ then A and B are equally likely.

Take a Bayesian model (X, Θ) with parameter space Π. We split the parameter space into two pieces: $\Pi = \Pi_0 \cup \Pi_1$ where $\Pi_0 \cap \Pi_1 = \emptyset$. We consider two competing hypothesis: H_0 is that the unknown parameter θ is within the set Π_0, and H_1 is that the unknown parameter θ is within the set Π_1.

Definition 7.1.2 The *prior odds* of H_0 against H_1 is defined to be

$$\frac{\mathbb{P}[\Theta \in \Pi_0]}{\mathbb{P}[\Theta \in \Pi_1]}.$$

Given the data x, the *posterior odds* of H_0 against H_1 is defined to be

$$\frac{\mathbb{P}[\Theta|_{\{X=x\}} \in \Pi_0]}{\mathbb{P}[\Theta|_{\{X=x\}} \in \Pi_1]}.$$

Note that the prior odds involves the prior Θ, and the posterior odds involve the posterior $\Theta|_{\{X=x\}}$, both otherwise the formulae are identical. We assume implicitly that $\mathbb{P}[\Theta \in \Pi_0]$ and $\mathbb{P}[\Theta \in \Pi_1]$ are both non-zero, which by Theorems 2.4.1 and 3.1.2 implies that the same is true for $\Theta|_{\{X=x\}}$. Note also that the prior and posterior odds are only well defined for proper prior and posterior distributions, or else we cannot make sense of the probabilities above.

It is often helpful to get a feel for how much the data has influenced the result of the test. For these purposes we also define the *Bayes factor*

$$B = \frac{\text{posterior odds}}{\text{prior odds}}. \tag{7.1}$$

Our next lemma shows why B is important. It is equal to the ratio of the likelihoods of the event $\{X = x\}$, i.e. of the data that we have, conditional on $\Theta \in \Pi_0$ and $\Theta \in \Pi_1$. In other words, B is the ratio of the likelihood of H_0 compared to H_1.

Lemma 7.1.3 *In the notation above, the Bayes factor satisfies* $B = \frac{L_{X|_{\{\Theta \in \Pi_0\}}}(x)}{L_{X|_{\{\Theta \in \Pi_1\}}}(x)}$ *where* L *denotes the likelihood function.*

Proof: We split the proof into two cases, depending on whether the Bayesian model is discrete or absolutely continuous. In the discrete case we have

$$B = \frac{\mathbb{P}[\Theta|_{\{X=x\}} \in \Pi_0]\mathbb{P}[\Theta \in H_1]}{\mathbb{P}[\Theta|_{\{X=x\}} \in \Pi_1]\mathbb{P}[\Theta \in H_0]} = \frac{\frac{\mathbb{P}[\Theta \in \Pi_0, X=x]}{\mathbb{P}[X=x]}\mathbb{P}[\Theta \in H_1]}{\frac{\mathbb{P}[\Theta \in \Pi_1, X=x]}{\mathbb{P}[X=x]}\mathbb{P}[\Theta \in H_0]} = \frac{\frac{\mathbb{P}[\Theta \in \Pi_0, X=x]}{\mathbb{P}[\theta \in \Pi_0]}}{\frac{\mathbb{P}[\Theta \in \Pi_1, X=x]}{\mathbb{P}[\theta \in \Pi_1]}} = \frac{\mathbb{P}[X|_{\{\Theta \in \Pi_0\}} = x]}{\mathbb{P}[X|_{\{\Theta \in \Pi_1\}} = x]}.$$

©Nic Freeman, University of Sheffield, 2024.

We have used equation (1.4) from Lemma 1.4.1 several times here. The continuous case is left for you, in Exercise **7.7** ■

As a rough guide to interpreting the Bayes factor, the following table[1] is often used:

Bayes factor	Interpretation: evidence in favour of H_0 over H_1
1 to 3.2	Indecisive / not worth more than a bare mention
3.2 to 10	Substantial
10 to 100	Strong
above 100	Decisive

Note that a high value of B only says that H_0 should be preferred over H_1. It does not tell us anything objective about how good our model (M_θ) is; it only tells us that $X|_{\{\Theta\in\Pi_0\}}$ is a better fit for x than $X|_{\{\Theta\in\Pi_1\}}$ is.

Values of the Bayes factor below 1 suggest evidence in favour of H_1 over H_0. In such a case we can swap the roles of H_0 and H_1, which corresponds to the Bayes factor changing from B to $1/B$, and we can then use the same table to discuss the weight of evidence in favour of H_1 over H_0.

[1] From Kass & Raftery (1995).

©Nic Freeman, University of Sheffield, 2024.

Example 7.1.4 Returning to Example 4.5.3, suppose that we wished to test the hypothesis that the speed camera is, on average, overestimating the speed to cars. According to our posterior, a car travelling at exactly 30mph will have a recorded speed with a $N(\mu, \frac{1}{\tau})$ distribution where $(\mu, \tau) \sim \text{NGamma}(30.14, 10.01, 6.00, 1.24)$.

The speed camera on average overestimates the speed when $\mu > 30$, and underestimates when $\mu < 30$. The probability that μ is exactly 30 is zero, because our posterior NGamma is a continuous distribution, so we will simply ignore that possibility. We don't care about the location of τ here so we simply allow it to take any value $\tau \in (0, \infty)$. This gives us hypothesis

$$\begin{aligned} H_0 &: \text{ that } (\mu, \tau) \subseteq \Pi_0 = (30, \infty) \times (0, \infty), \\ H_1 &: \text{ that } (\mu, \tau) \subseteq \Pi_1 = (-\infty, 30) \times (0, \infty). \end{aligned}$$

We have

$$\mathbb{P}[(\mu, \tau) \in \Pi_0] = \int_{30}^{\infty} \int_{0}^{\infty} f_{\text{NGamma}(30.14,10.01,6.00,1.24)}(\mu, \tau)\, d\tau\, d\mu = 0.82,$$

computed numerically and rounded to two decimal places. Note that $\mathbb{P}[(\mu, \tau) \in \Pi_1] = 1 - \mathbb{P}[(\mu, \tau) \in \Pi_0]$, which gives a posterior odds ratio of

$$\frac{\mathbb{P}[\Theta|_{\{X=x\}} \in H_0]}{\mathbb{P}[\Theta|_{\{X=x\}} \in H_1]} = \frac{0.82}{1 - 0.82} = 4.56$$

again rounded to two decimal places. The prior odds ratio, calculated via the same procedure, is exactly 1. This occurs because of the symmetry of the prior $\text{NGamma}(30, \frac{1}{10^2}, 1, \frac{1}{5})$ distribution (this symmetry is visible in the sketch in Example 4.5.3) gives that $\mathbb{P}[\text{NGamma}(30, \frac{1}{10^2}, 1, \frac{1}{5}) \in \Pi_1] = \mathbb{P}[\text{NGamma}(30, \frac{1}{10^2}, 1, \frac{1}{5}) \in \Pi_0] = \frac{1}{2}$. Hence the Bayes factor for this hypothesis test is

$$B = \frac{4.56}{1.00} = 4.56. \tag{7.2}$$

Based on our table above, we have substantial evidence that the speed camera is overestimating speeds.

A potential problem with our test is that we have not cared about *how much* the camera is overestimating speeds. The (marginal) mean of μ in our posterior distribution is 30.11, which is only slightly larger than the true speed 30, and this suggests that the error is fairly small. We would need to be careful about communicating the result of our test, to avoid giving the wrong impression.

Note that we have used a small amount of Bayesian shorthand in this example, by writing μ and τ for both random variables and samples of these random variables.

©Nic Freeman, University of Sheffield, 2024.

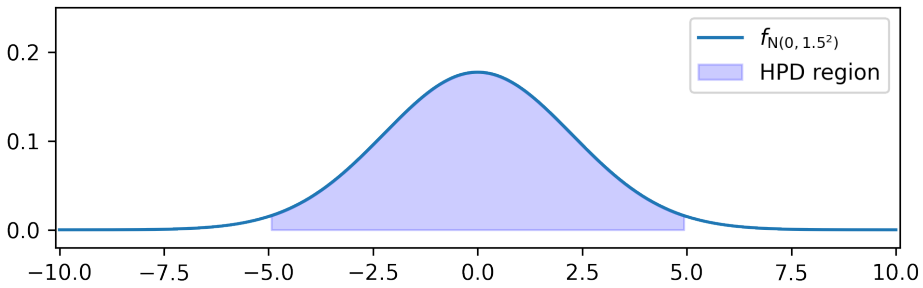
7.2 High posterior density regions

In this section we look at ways to report interval estimates for the unknown parameter θ. The key term used in the Bayesian framework is the following definition.

Definition 7.2.1 A *high posterior density region* is a subset $\Pi_0 \subseteq \Pi$ that is chosen to minimize the size of Π_0 and maximize $\mathbb{P}[\Theta|_{\{X=x\}} \in \Pi_0]$.

This is a practical definition rather than a mathematical one: we must choose how to balance the minimization of Π_0 against maximization of $\mathbb{P}[\Theta|_{\{X=x\}} \in \Pi_0]$ as best we can, and in some situations there is no single right answer. We write HPD region, for short. In some textbooks they are known as HDRs.

Example 7.2.2 Suppose that our posterior has come out as $\Theta|_{\{X=x\}} \sim \mathrm{N}(0, 1.5^2)$.



We choose our HPD region to be $[-5, 5]$. This region is much smaller than the range of Θ, which is the whole of $\mathbb{R}$. The probability that it contains $\Theta|_{\{X=x\}}$ is given by $\mathbb{P}[-5 \leq N(0, 1.5^2) \leq 5] = 0.97$ to 2 decimal places.

More generally, if we are dealing with a continuous distribution with a single peak then it is common to choose a HPD region of the form $\Pi_0 = [a, b]$ where

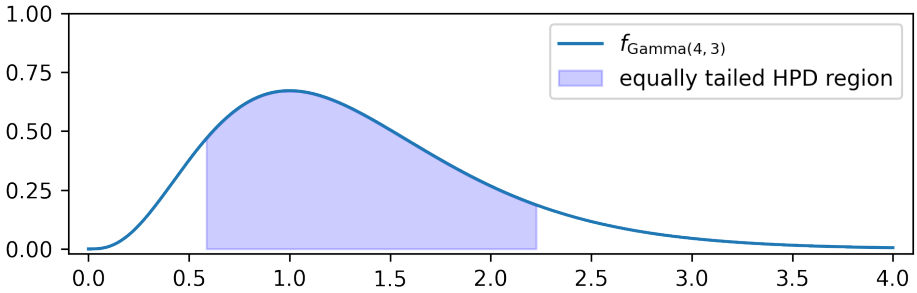
$$\mathbb{P}\left[\Theta|_{\{X=x\}} < a\right] = \mathbb{P}\left[\Theta|_{\{X=x\}} > b\right] = \frac{1-p}{2} \tag{7.3}$$

and some value is picked for $p \in (0, 1)$. HPD intervals of this type are said to be *equally tailed*. They always contain the mode of $\Theta|_{\{X=x\}}$, and from (7.3) we have $\mathbb{P}[\Theta|_{\{X=x\}} \in [a, b]] = p$. By symmetry the HPD region in Example 7.2.2 is equally tailed, and we will give an asymmetric case in Example 7.2.3.

A value of p close to 1 gives a wide interval and a high value for $\mathbb{P}[\Theta|_{\{X=x\}} \in [a, b]]$, whilst a value of p close to 0 gives a thinner interval but a lower value for $\mathbb{P}[\Theta|_{\{X=x\}} \in [a, b]]$. As in Example 7.2.2, we want to choose $p \in (0, \frac{1}{2})$ to make $[a, b]$ thin *and* make $\mathbb{P}[\Theta|_{\{X=x\}} \in [a, b]]$ large, if possible.

©Nic Freeman, University of Sheffield, 2024.

Example 7.2.3 Suppose that our posterior has come out as $\Theta|_{\{X=x\}} \sim \Gamma(3,4)$. We want an equally tailed HPD region $[a,b]$ such that $\mathbb{P}[\Theta|_{\{X=x\}} \in [a,b]] = 0.8$.



The region is chosen by finding a such that $\mathbb{P}[\Theta|_{\{X=x\}} < 1] = \frac{1-0.8}{2} = 0.1$ and b such that $\mathbb{P}[\Theta|_{\{X=x\}} > b] = \frac{1-0.8}{2} = 0.1$. These values were found numerically to be $a = 0.58$ and $b = 2.23$, to two decimal places. You can find the code that generated this example as part of Exercise 7.2.3.

Remark 7.2.4 (⊘) It is possible to construct a form of hypothesis testing based on HPDs. For example, we might by choose an HPD Π_0 with a 95% probability of containing $\Theta|_{\{X=x\}}$, and then we then reject the hypothesis $\theta = \theta_0$ if and only if $\theta_0 \notin \Pi_0$. This approach is known as *Lindley's method*.

Remark 7.2.5 (⊘) We can define *high prior density regions* in the same ways as detailed above, with Θ in place of $\Theta|_{\{X=x\}}$. These are less useful for parameter estimation although they can be useful for prior elicitation. We implicitly used equally tailed regions of this type in Example 5.1.1, when we asked an elicitee to estimate their 25th and 75th percentiles.

©Nic Freeman, University of Sheffield, 2024.

7.3 Point estimates

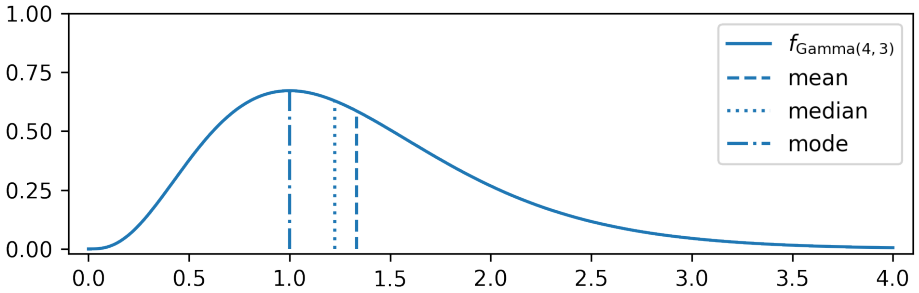
If, for some reason, it is necessary to give a point estimate θ_0 for the unknown parameter θ, then we can obtain one in various different ways from the posterior $\Theta|_{\{X=x\}}$. Common choices are:

1. The *mean* $\theta_0 = \mathbb{E}[\Theta|_{\{X=x\}}]$.
2. The *median* θ_0 such that $\mathbb{P}[\Theta|_{\{X=x\}} \leq \theta_0] = \mathbb{P}[\Theta|_{\{X=x\}} \geq \theta_0] = \frac{1}{2}$.
3. The *mode* $\theta_0 = \arg\max_{\theta\in\Pi} L_{\Theta|_{\{X=x\}}}(\theta)$.

When doing so, we should be wary that θ_0 contains much less information than the full posterior distribution $\Theta|_{\{X=x\}}$. If Θ is close to θ_0 with high probability then we can hope it provides a reasonable approximation, but there is no guarantee in any case, and we should plot the distribution of $\Theta|_{\{X=x\}}$ (and, ideally, the corresponding sampling distributions) to assess this.

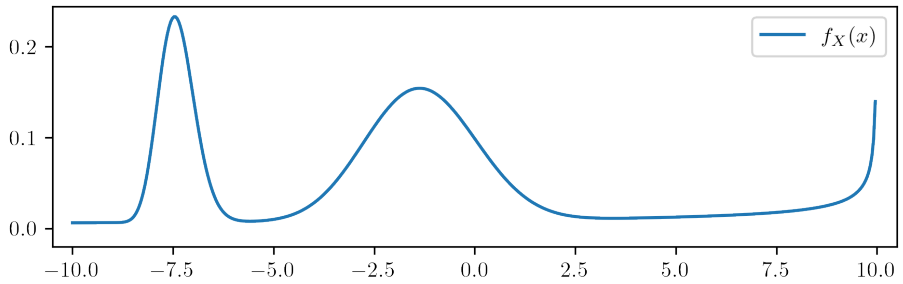
Note that some or all of these point estimates may fail to be well defined, dependent upon the distribution of X. For example the Cauchy distributions does not have a well-defined mean. Formulae for means and variances are listed on the reference sheets in Appendix A; modes can usually be obtained (for continuous distribution) via differentiation; medians are usually obtained either via symmetry or numerically.

Example 7.3.1 In Example 7.2.2 the mean, median and mode of $\Theta|_{\{X=x\}} \sim \mathrm{N}(0, 1.5^2)$ are all equal to zero, due to the symmetry of the normal distribution. In Example7.2.3, in which $\Theta|_{\{X=x\}} \sim \mathrm{Gamma}(3, 4)$, they are



All of these are reasonable point estimates for $\mathrm{Gamma}(0, 1.5^2)$. The distribution of $\mathrm{Gamma}(0, 1.5^2)$ spreads out across a range of parameters, and we should make this clear in our analysis, perhaps by giving an HPD interval alongside the point estimate.

Lastly, consider if we had obtained a posterior distribution with the p.d.f. from Exercise 1.1:



There is no reasonable way to summarize this distribution with a point estimate. In a case like this we should decline to give a point estimate, even if we are asked for one.

©Nic Freeman, University of Sheffield, 2024.

7.4 Comparison to classical methods

You will have seen a different method of carrying out a hypothesis test before, looking something like this.

Definition 7.4.1 The *classical hypothesis test* is the following procedure:

1. Choose a model family $(M_\theta)_{\theta\in\Pi}$, choose a value $\theta_0 \in \Pi$ and define H_0 to be the model M_{θ_0}. This is often written in shorthand as $H_0 : \theta = \theta_0$.

2. Calculate a value p as follows. Assume that H_0 is true i.e. use the model M_{θ_0} and using this model, calculate p to be the probability of observing data that is (in some chosen sense) 'at least as extreme' as the data x that we actually observed.

   If p is sufficiently small (in some chosen sense) then reject H_0.

There is no need for an 'alternative hypothesis' here. More specifically, rejecting H_0 means that we think it is unlikely that our chosen model M_{θ_0} would generate the data x, so consequently we think it is unlikely that M_{θ_0} is a good model.

There is nothing else to say here! Rejecting H_0 does not mean that the 'alternative hypothesis' H_1 that $\theta \neq \theta_0$ is accepted (or true). If p turns out to be small it means that either (i) M_{θ_0} is a good model and our data x was unlikely to have occurred or, (ii) M_{θ_0} is a bad model for our data. Neither statement tells us what a good model might look like. Unfortunately hypothesis testing is very often misunderstood, and rejection of H_0 is incorrectly treated as though it implies that H_1 is true.

If we do not reject H_0, then it means that the model M_{θ_0} is reasonably likely to generate the data we have. This leaves open the possibility that there may be lots of other models, not necessarily within our chosen model family, that are also reasonably likely to generate the data we have. This point is sometimes misunderstood too.

Example 7.4.2 A famous example of these mistakes comes from the 'clever Hans' effect. Hans was a horse who appeared to be able to do arithmetic, owned by a mathematics teacher Wilhelm von Osten. Von Osten would ask Hans (by speaking out loud) to answer to various questions and Hans would reply by tapping his hoof. The number of taps was interpreted as a numerical answer. Hans answered the vast majority of questions correctly.

To construct a hypothesis test using Definition 7.4.1, take a model family $M_\theta \sim \text{Bernoulli}(\theta)^{\otimes n}$, where the data $x = (x_1, \ldots, x_n)$ corresponds to $x_i = 1$ for solving the n^{th} question correctly, and $x_i = 0$ for incorrectly. Let us suppose that the probability of Hans solving a question correctly by guessing at random is $\theta = \frac{1}{2}$. So, take H_0 to be $\theta = \frac{1}{2}$, and the model we wish to test is $M_{\frac{1}{2}}$. The horse is asked $n = 10$ questions, and it answers them all correctly. Our model $M_{\frac{1}{2}}$ says the probability of this is $(\frac{1}{2})^{10} \approx 0.001 = p$. We reject H_0. Taking any value $\theta \leq \frac{1}{2}$ will lead to the same conclusion.

So, we expect that our model M_θ is a bad description of reality, for each $\theta \leq \frac{1}{2}$. This does not mean that we must accept H_1 and believe the horse is doing arithmetic i.e. that some alternative model M_θ is correct for some larger value of θ. In fact, what is going on here is that Hans has learnt to read the body language of Wilhelm von Osten, who leans in forwards whilst Hans is tapping his hoof and leans back upright as soon as the correct number of taps has been reached.

©Nic Freeman, University of Sheffield, 2024.

This was established by the psychologist Oskar Pfungst, who tested Hans and von Osten under several different conditions in a laboratory.

In short, our model that the horse 'solves' questions is wrong. The horse *answers* questions correctly but it does not *solve* questions. To distinguish between these two situations we need a better model than (M_θ), as Pfungst did in his laboratory. After the investigations by Pfungst were done, von Osten refused to believe what Pfungst had discovered, and continued to show Hans around Germany. They attracted large and enthusiastic crowds, and made a substantial amount of money from doing so – many in his audience wondered if they should accept H_1.

The point of including this example in our course – in which we focus on Bayesian methods – is to note that errors in interpretation are less common when using Bayesian approaches. The reason is simply that the Bayesian approach uses the framework of conditional probability, so we state our results in terms of conditional probabilities and odds ratios. This makes our assumptions and conclusions clear.

By contrast, the p-value from Definition 7.4.1 is not a conditional probability because conditioning the model (M_θ) on the event $\{\Theta = \theta_0\}$ is only possible if we have a random variable Θ that we can condition on the event $\{\Theta = \theta_0\}$, and Definition 7.4.1 does not include this random variable. However, if we take Θ to be a uniform prior and (M_θ) is well-behaved enough (e.g. Assumption 1.3.2) then $M_\Theta|_{\{\Theta=\theta_0\}} \overset{\mathrm{d}}{=} M_{\theta_0}$ can be shown. In that situation the p-value is the conditional probability, given $\{\Theta = \theta_0\}$, of observing data that is (in some chosen sense) at least as extreme as what we did observe. Calculating p is often difficult and instead it is common to use approximation theorems. These theorems[2] tend to contain complicated assumptions that are difficult to state correctly, particularly when data points may not be fully independent of each other. The combined result is that Definition 7.4.1 gives a procedure with many potential sources of error.

Remark 7.4.3 I do not mean to imply that using the Bayesian framework would certainly have avoided making the mistake detailed in Example 7.4.2. Only that, because the framework would make us state our assumptions and conclusions more clearly, we can then more easily question which of our assumptions was incorrect.

When a horse claims to do arithmetic we are naturally suspicious. In more subtle situations it is harder to find mistakes.

Similar considerations apply to the comparison between HPDs and confidence intervals. For example, a 95% confidence $[a, b]$ intervals is often incorrectly treated as a statement that the 'event' $\theta \in [a, b]$ has 95% probability. An HPD actually *is* a statement to that effect, about the posterior distribution $\Theta|_{\{X=x\}}$, which makes it much easier to interpret.

[2] (⊘) E.g. Central limit theorems, Wilks' theorem.

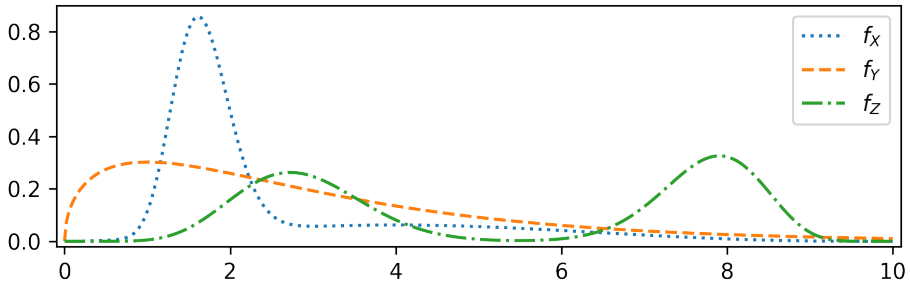
©Nic Freeman, University of Sheffield, 2024.

Figure 7.1: Hans the horse in 1904, correctly answering arithmetic questions set by his owner Wilhem von Osten.

©Nic Freeman, University of Sheffield, 2024.

7.5 Exercises on Chapter 7

⋆ **7.1** Consider the following sketch of three probability density functions. In each case, discuss whether you think it would reasonable to approximate the distribution with a point estimate. If so, would you prefer the mean, median or mode?



⋆⋆ **7.2** In the situation of Example 2.3.3, the experiments test the success of a medical treatment. Each trial is expensive, so it is decided by the regulator that further trials of the treatment will be carried out if and only if there is substantial evidence that $\theta > 0.2$.

(a) Using the original prior $\text{Beta}(2, 8)$ and the posterior $\text{Beta}(11, 19)$ odds obtained after the second round of trials, find the prior and posterior odds ratios of $\theta > 0.2$, and find the associated Bayes factor.

(b) Show that the reference prior for the Binomial distribution is $\Theta \sim \text{Beta}(\frac{1}{2}, \frac{1}{2})$ (this generalizes Example 5.3.4).

How much does using this prior change the results of your analysis in (a)?

(c) Discuss briefly whether the regulator would be interested to see the results in (a), or (b), or both, when making their decision as to whether the trials should proceed further.

⋆⋆ **7.3** Inside the file `7_earthquakes_japan.csv` you will find a dataset listing the number of earthquakes, of magnitude 7.5 or higher, that occurred in Japan during the years 1984-2023. We model the occurrence of earthquakes in a particular year as Poisson(λ), independently for each year, and we consider the hypothesis $H_0 : \lambda \geq 2$.

(a) In Exercise **5.3** we found that the reference prior for the Poisson distribution was

$$f(\lambda) \propto \begin{cases} \lambda^{-1/2} & \text{for } \lambda > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Are the prior odds for H_0 defined, with respect to this prior?

(b) Recall from Exercise **4.3** that the Gamma distribution provides a conjugate prior to the Poisson model family. Using the data given and the weakly informative prior $\lambda \sim \text{Gamma}(\frac{5}{4}, \frac{1}{5})$, find the posterior distribution for λ, and find the prior and posterior odds ratios of the evidence in favour of H_0 over H_1. Calculate the associated Bayes factor and comment on the evidence for H_0.

(c) Repeat part (b) for $H_0 : \lambda \geq 3$.

©Nic Freeman, University of Sheffield, 2024.

⋆ **7.4** Inside the files `7_hpd_example.ipynb` and `7_hpd_example.Rmd` you will find code to generate the figure in Exercise 7.2.3. Modify this code (using either file) to construct an equally tailed HPD region $[a, b]$ for the ChiSquared(3) distribution, such that $\mathbb{P}[\text{ChiSquared}(3) \in [a, b]] = 0.7$.

⋆⋆ **7.5** In this question we model the number of hurricanes per year making landfall in the United States of America using the Poisson(θ) distribution. We will assume that each year is independent and we will use a prior $\theta \sim \text{Exp}(\lambda)$.

(a) Perform an elicitation method of your choice from Section 5.1, either alone or with a parter, to choose the value of λ that best expresses your prior beliefs about the number of hurricanes per year that make landfall in the USA.

(b) Inside the file `7_hurricane_landfalls_usa.csv` you will find a dataset corresponding to the years 2015-2022. Use the result of Exercise **4.3** to find the posterior distribution resulting from your prior and this dataset.

Hint: Recall that $\text{Exp}(\lambda) \sim \text{Gamma}(1, \lambda)$.

(c) Sketch an equally tailed 95% HPD region for θ.

7.6 In this question we consider HPD intervals for the mean, using the model $N(\mu, \phi)^{\otimes n}$ family, with both parameters unknown. We will use the reference prior for this family, which is an improper prior with density function

$$f(\mu, \phi) = \begin{cases} \frac{1}{\phi} & \text{for } \tau > 0 \text{ and } \mu \in \mathbb{R} \\ 0 & \text{otherwise.} \end{cases}.$$

Write $\bar{x} = \sum_1^n x_i$ and $S^2 = \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2$. It is known that, for this model and reference prior, the variable $t|x = \frac{(\mu|x) - \bar{x}}{S/\sqrt{n}}$ has distribution $t|x \sim \text{Student-t}(n-1)$.

⋆⋆ (a) Inside the file `7_wheat_yields_uk.csv` you will find a dataset of the UK wheat yields, measured in tonnes per hectare of farmland, for the years 1983-2022. You may assume that the model $N(\mu, \phi)^{\otimes 40}$ is appropriate for this data. Construct and sketch an equally tailed 95% HPD region for t, and hence give a 95% HPD region for μ.

⋆⋆⋆ (b) Find the posterior density function $f(\mu, \phi)$, for the model $N(\mu, \phi)^{\otimes n}$ with the given reference prior. Hence, verify that the distribution of $t|x$ is indeed Student-t$(n-1)$.

⋆⋆⋆ **7.7** Prove the continuous case of Lemma 7.1.3. You should use Theorem 3.1.2 along with Exercises **1.8** and **3.6** to justify your calculations.

©Nic Freeman, University of Sheffield, 2024.

Chapter 8

Computational methods

We noted at several points that conjugate pairs, from Chapter 4, do not provide enough flexibility for many practical situations. Instead, Bayesian statistics is heavily reliant on a family of computational techniques, introduced within this chapter. They generate samples from the posterior distribution and have a moderate computational cost. With simple model families of the kind we have worked with throughout the course it is reasonable to use desktop machines. More complex model families can require larger machines.

Throughout this chapter we assume the same setup as in Chapter 7, which we repeat here for convenience. We work with a discrete or absolutely continuous Bayesian model (X, Θ), where we have data x and posterior $\Theta|_{\{X=x\}}$. We keep all of our usual notation: the parameter space is Π, the model family is $(M_\theta)_{\theta\in\Pi}$, and the range of the model is R. Note that M_θ could have the form $M_\theta \sim (Y_\theta)^{\otimes n}$ for some random variable Y_θ with parameter θ, corresponding to n i.i.d. data points.

©Nic Freeman, University of Sheffield, 2024.

8.1 Approximate Bayesian computation (⊘)

In this section we describe a numerical method for calculating the posterior $\Theta|_{\{X=x\}}$ that is based on rejection sampling. Recall that we used rejection sampling in Section 1.4 to prove Lemma 1.4.1, and also to give some intuition for our first examples of conditioning.

The algorithm we study here is known as Approximate Bayesian Computation, or ABC for short. We will describe it first for discrete data, in the situation where the prior (and, consequently, the posterior) are also discrete distributions. We haven't studied this case in any of our previous chapters, so let us first introduce it here.

Definition 8.1.1 (Bayesian model with discrete parameters and discrete data) Take a prior with p.m.f. $p_\Theta(\theta)$ and a discrete model family $(M_\theta)_{\theta\in\Pi}$ where Π is a finite or countable set. The Bayesian model (X, Θ) has the law

$$\mathbb{P}[X = x, \Theta = \theta] = \mathbb{P}[M_\theta = x]p_\Theta(\theta).$$

It is straightforward to sum over x and obtain the prior distribution $\mathbb{P}[\Theta = \theta] = p_\Theta(\theta)$, and also to sum over θ and obtain the sampling distribution $\mathbb{P}[X = x] = \sum_{\theta\in\Pi}\mathbb{P}[M_\theta = x]p_\Theta(\theta)$. For $x \in R_X$ we have $\mathbb{P}[X = x] > 0$ and thus the posterior $\Theta|_{\{X=x\}}$ is defined via Lemma 1.4.1. Also using Lemma 1.4.1, the conditional distribution $X|_{\{\Theta=\theta\}}$ satisfies

$$\mathbb{P}[X|_{\{\Theta=\theta\}} = x] = \frac{\mathbb{P}[X = x, \Theta = \theta]}{\mathbb{P}[\Theta = \theta]} = \frac{\mathbb{P}[M_\theta = x]p_\Theta(\theta)}{p_\Theta(\theta)} = \mathbb{P}[M_\theta = x]$$

so $X|_{\{\Theta=\theta\}} \overset{\mathrm{d}}{=} M_\theta$.

In the context of Definition 8.1.1 the *ABC algorithm* for generating samples from $\Theta|_{\{X=x\}}$ is the following:

1. Sample θ_0 from the discrete distribution Θ.
2. Sample x_0 from the discrete distribution $M_\theta \overset{\mathrm{d}}{=} X|_{\{\Theta=\theta\}}$.
3. Then:
   - if $x \neq x_0$, go back to step one;
   - if $x = x_0$,accept θ_0 as a sample of $\Theta|_{\{X=x\}}$.

This algorithm is precisely the strategy of our proof for Lemma 1.4.1, written as an algorithm and adapted to the special case of Definition 8.1.1. It generates a single sample of the distribution $\Theta|_{\{X=x\}}$. We can run the algorithm again to obtain more samples.

The ABC algorithm outlined above requires only that we have the ability to take samples from discrete distributions with known probability mass functions. To handle cases with continuous priors and/or data, we also need to be able to sample from continuous distributions with known probability density functions. The modifications are as follows:

- If Θ is continuous and (M_θ) is discrete then we can adopt the same algorithm, with the modification that in step 1 we must now sample from a continuous distribution rather than a discrete distribution.

©Nic Freeman, University of Sheffield, 2024.

- If (M_θ) is continuous then in step 3 we will have $\mathbb{P}[x = x_0] = 0$. In this case the simplest strategy is to fix some $\epsilon > 0$ and accept θ_0 as an *approximate* sample of $\Theta|_{\{X=x\}}$ if $|x - x_0| \leq \epsilon$.

  This idea is based on (1.9), which stated that if $\Theta|_{\{X=x\}}$ was to be defined then it should be defined to be the limit as $\epsilon \to 0$ of $\Theta|_{\{|X-x|\leq\epsilon\}}$. The terminology 'Approximate' Bayesian Computation comes from this step.

  More complex strategies for comparing x and x_0 can also be used, with the aim of focusing on the aspects of the data that are most important to us.

The ABC algorithm as described above has a serious drawback. In discrete cases the probability that $x = x_0$ (in step 3) can be extremely small, meaning that we have to go around the loop $1 \to 2 \to 3 \to 1$ many times before we find a sample we can accept. In continuous cases, choosing ϵ close to 0 obtains results in good approximation but introduces the same problem; that the probability of accepting the same x_0 becomes small. To get handle this difficulty, various ways of sampling x_0 (in step 2) have been developed to increase the acceptance probability, without changing the distribution of x_0. One such method is ABC-MCMC, which uses ideas from Section 8.3 to sample x_0. Another method is *sequential* ABC, where x_0 is sampled as a perturbation in a carefully chosen direction from the (rejected) x_0 in the previous iteration of the loop. We will not detail such methods here. They are very popular in some applications.

©Nic Freeman, University of Sheffield, 2024.

8.2 Metropolis-Hastings

In order to perform Bayesian inference computationally, the main requirement is that we can obtain samples from the posterior distribution $\Theta|_{\{X=x\}}$. We know the p.d.f. from Theorem 2.4.1/3.1.2, but this The most popular strategy for doing so is based on the Metropolis-Hastings algorithm. We will describe the Metropolis-Hastings algorithm in this section, and explain its application to Bayesian inference in Section 8.3.

8.2.1 The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is a general technique for producing samples from a distribution. We will describe it in the case where we take samples of a continuous random variable Y with p.d.f. f_Y and range $R_Y \subseteq \mathbb{R}^d$. The key ingredient of the algorithm is a joint distribution (Y, Q), where $Q|_{\{Y=y\}}$ and $Y|_{\{Q=y\}}$ are both well defined for all $y \in R_Y$, both with the same range as Y.

Example 8.2.1 Given some continuous distribution Y with range $\mathbb{R}$, a common choice is to take $Q = Y + \mathrm{N}(0, \sigma^2)$ where $\sigma > 0$ is a constant.

The *Metropolis-Hastings algorithm* is the following. For now, it won't be obvious why this algorithm generates samples of Y, but we will address this point soon after. Let y_0 be a point within R_Y. Then, given y_m we define y_{m+1} as follows.

1. Generate a *proposal point* $\tilde{y}$ from the distribution of $Q|_{\{Y=y_m\}}$.
2. Calculate the value of
$$\alpha = \min\left\{1, \frac{f_{Y|_{\{Q=\tilde{y}\}}}(y_m) f_Y(\tilde{y})}{f_{Q|_{\{Y=y_m\}}}(\tilde{y}) f_Y(y_m)}\right\} \tag{8.1}$$
3. Then, set
$$y_{m+1} = \begin{cases} \tilde{y} & \text{with probability } \alpha, \\ y_m & \text{with probability } 1-\alpha. \end{cases} \tag{8.2}$$

The distribution $Q|_{\{Y=y\}}$ is called the *proposal* distribution, based on its role in steps 1 and 2. The two cases in step 3 are usually referred to as *acceptance* (when $y_{m+1} = \tilde{y}$) and *rejection* (when $y_{m+1} = y_m$). The key point is that the algorithm only needs samples from the proposal distribution; we can run it without needing to sample of the distribution of Y directly! The Metropolis-Hastings algorithm is useful in cases where the distribution of Y is unknown or is too complicated to efficiently sample from.

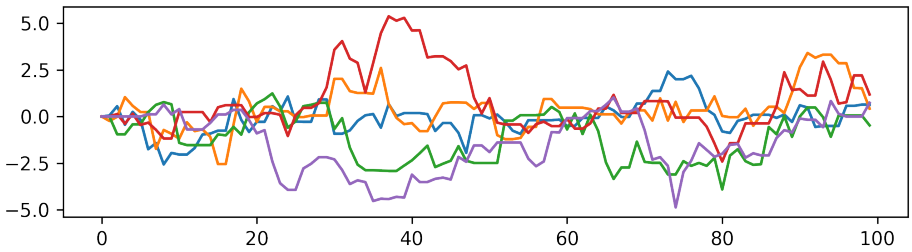
Theorem 8.2.2 *Let (y_m) be the random sequence obtained from the Metropolis-Hastings algorithm. Then for all $A \subseteq R_Y$ we have $\mathbb{P}[y_m \in A] \to \mathbb{P}[Y \in A]$ as $m \to \infty$.*

Theorem 8.2.2 says that if we run the MH algorithm for a long time, so that m becomes large, the random value of y_m will have a similar distribution to Y. We won't be able to prove Theorem 8.2.2 in this course but we will give a detailed idea of why it is true in Section 8.2.3 (which is off-syllabus). As you might expect, this will involve the precise form of (8.1).

From Lemma 1.6.1 we can rewrite equation (8.1) as $\alpha = \min\left(1, \frac{f_{Y,Q}(\tilde{y}, y_m)}{f_{Y,Q}(y_m, \tilde{y})}\right)$. This is a nicer formula, but the convention that you will find in all textbooks is to write the form (8.1). The reason for this will become clear in Section 8.2.2.

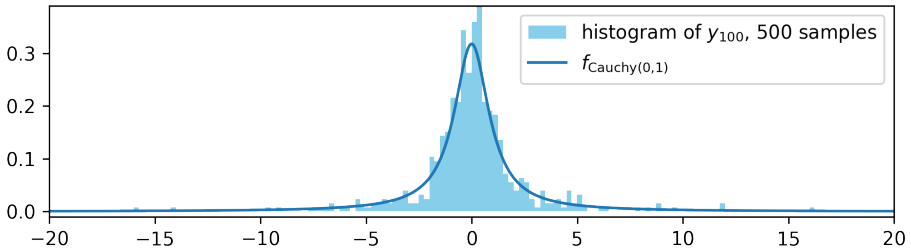
©Nic Freeman, University of Sheffield, 2024.

Example 8.2.3 Here's some examples of the random sequence (y_m) generated by the MH algorithm, in the case $Y \sim \text{Cauchy}(0,1)$ with $Q = Y + \text{N}(0,1)$ as in Example 8.2.1.

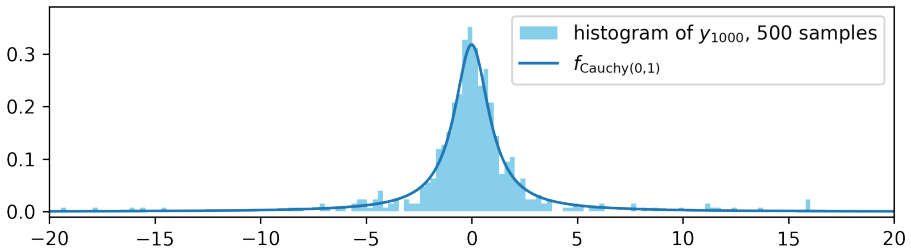


We've shown five sample runs of $(y_m)_{m=1}^{100}$, starting from zero in each case. You can see that sometimes for a few steps of time passes whilst a path does not move – that is when the proposal $\widetilde{y}$ is rejected a few times in a row. When the paths do move, each movement is an (independent) $\text{N}(0,1)$ random variable.

Next we run the MH algorithm 500 times, and in each case we record the value of y_{100}. Theorem 8.2.2 tells us that each y_{100} should be approximately a sample of $\text{Cauchy}(0,1)$, so by taking 500 samples we should be able to see the shape of the distribution. We plot these values as a histogram and compare to the p.d.f. of $\text{Cauchy}(0,1)$, giving



We can see that the histogram is a reasonable match for $f_{\text{Cauchy}(0,1)}$, so the MH algorithm is behaving as Theorem 8.2.2 predicts. If we let the MH algorithm have more steps, so that we consider y_{1000} instead of y_{100}, then we obtain a better approximation:



Of course, we could also obtain a better approximation to the distribution of $Y \sim \text{Cauchy}(0,1)$ by taking more samples. The code that generated the plots above is given to you for use in several of the exercises at the end of this chapter.

In statistics you will often hear the terminology 'y_m has converged' used to mean that m is large enough that y_m has approximately the same distribution as Y. This is a misuse of terminology, but it is common and quite helpful in practice. The period before is sometimes known as 'burn in'.

©Nic Freeman, University of Sheffield, 2024.

8.2.2 The Metropolis algorithm

A common technique when using the MH algorithm is to choose Q in such a way that

$$f_{Q|_{\{Y=y\}}}(\tilde{y}) = f_{Y|_{\{Q=\tilde{y}\}}}(y) \tag{8.3}$$

for all y and $\tilde{y}$. The point of doing so is that it greatly simplifies the formula (8.1) for α, because the terms on top and bottom involving conditional densities then cancel. The algorithm for updating y_n then becomes:

1. Generate a *candidate point* $\tilde{y}$ from the distribution of $Q|_{\{Y=y_n\}}$.
2. Calculate the value of
$$\alpha = \min\left\{1,\ \frac{f_Y(\tilde{y})}{f_Y(y_n)}\right\} \tag{8.4}$$
3. Then, set
$$y_{n+1} = \begin{cases} \tilde{y} & \text{with probability } \alpha, \\ y_n & \text{with probability } 1-\alpha. \end{cases}$$

This is known as the *Metropolis* algorithm, or sometimes as the *symmetric Metropolis-Hastings* algorithm.

Example 8.2.4 It is often straightforward to write down a Q such that (8.3) holds. For example, Suppose that Y is any random variable with range $\mathbb{R}$ and take $Q = Y + \mathrm{N}(0, \sigma^2)$ as in Example 8.2.1. Then from Remark 1.6.4 we have

$$Q|_{\{Y=y\}} \stackrel{\mathrm{d}}{=} y + \mathrm{N}(0,\sigma^2) \stackrel{\mathrm{d}}{=} \mathrm{N}(y,\sigma^2),$$
$$Y|_{\{Q=\tilde{y}\}} \stackrel{\mathrm{d}}{=} \tilde{y} - \mathrm{N}(0,\sigma^2) \stackrel{\mathrm{d}}{=} \mathrm{N}(\tilde{y},\sigma^2).$$

Hence

$$f_{Q|_{\{Y=y\}}}(\tilde{y}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\tilde{y})^2}{2\sigma^2}} = f_{Y|_{\{Q=\tilde{y}\}}}(y).$$

More generally, if Y has range $\mathbb{R}^d$ then (8.3) will hold whenever $Q = Y + Z$ where Z is a continuous random variable with a distribution that is symmetric about zero i.e. $f_Z(z) = f_Z(-z)$. Proving this fact is Exercise **??**.

This case is known as the *random-walk Metropolis* algorithm.

©Nic Freeman, University of Sheffield, 2024.

8.2.3 Why does Metropolis-Hastings work? (⊘)

In order to explain why the MH algorithm for (Y, Q) generates samples of Y, we will need some of the terminology that has been introduced in earlier courses on Markov chains. These courses are recommended pre-requisites to this course, but they are not compulsory pre-requisites, so this section is off-syllabus. We will sketch out parts of it in lectures, if there is time.

The sequence (y_m) defined by the MH algorithm in Section 8.2.1 is an example of a Markov chain with state space $R = R_Y$, the range of Y. In earlier courses you have studied Markov chains with discrete (i.e. finite or countable) state spaces, but in this case the state space is uncountable, because Y is a continuous random variable. Processes of this type are known as Markov chains in *continuous space*.

The key ingredients of a Markov chain with a finite or countable state space are its transition probabilities, which record the probabilities of moving between various states. In continuous space the equivalent concept is the function

$$p(x, A) = \mathbb{P}\left[X_{m+1}|_{\{X_m=x\}} \in A\right], \tag{8.5}$$

which is known as a *transition function* for the Markov chain (X_m). It gives the probability of moving to a state within A, from state x, where $A \subseteq R$ and R is the state space of the chain. You will sometimes see the right hand side of this equation written as $\mathbb{P}[X_{m+1} \in A | X_m = x]$, with the same meaning.

We say that a continuous random variable X with p.d.f. $f_X(x)$ is a *stationary distribution* for the chain (X_m) if when $X_m \overset{\mathrm{d}}{=} X$ we have also that $X_{m+1} \overset{\mathrm{d}}{=} X$. In symbols, this requirement means that $\mathbb{P}[X \in A] = \mathbb{P}[X_{m+1}|_{\{X_m=X\}} \in A]$, or equivalently

$$\mathbb{P}[X \in A] = \int_R p(x, A) f_X(x)\, dx \tag{8.6}$$

for all $A \subseteq R$. The expression on the right hand side here comes from (8.5), using that $\mathbb{P}[X_{m+1}|_{\{X_m=X\}} \in A] = \mathbb{E}[p(X, A)]$.

The definitions of periodicity, irreducibility and the various types of recurrence can be upgraded into continuous space in a natural way. Moreover, there is a convergence theorem for Markov chains in discrete space, which gives conditions (similar to those for discrete space) for the chain to converge to a unique stationary distribution, as time becomes large. We will not cover these ideas here, but note that our condition in Section 8.2.1 that $Q|_{\{Y=y\}}$ is a continuous random variable with range R_Y means that the sequence (y_m) might jump to anywhere within R_Y on any step of time. Under that condition the convergence theorem applies and it is known that the chain (y_m) will converge to a unique stationary distribution. The stationary distribution will be a continuous random variable and will satisfy (8.6).

We will show here that the transition function given by the MH algorithm satisfies (8.6) with stationary distribution Y. When this fact is combined with the convergence theorem for continuous space Markov chains, it leads to Theorem 8.2.2 – but we will only cover the calculation of the stationary distribution here. The transition function given by the MH algorithm is

$$\begin{aligned} p(x, A) &= \mathbb{P}\left[\text{Bernoulli}(\alpha_{x,Q|_{\{Y=x\}}}) = 1 \text{ and } Q|_{\{Y=x\}} \in A\right] + \mathbb{1}_{\{x\in A\}}\mathbb{P}\left[\text{Bernoulli}(\alpha_{x,Q|_{\{Y=x\}}}) = 0\right] \\ &= \mathbb{E}\left[\text{Bernoulli}(\alpha_{x,Q|_{\{Y=x\}}})\mathbb{1}_{\{Q|_{\{Y=x\}}\in A\}}\right] + \mathbb{1}_{\{x\in A\}}\mathbb{E}\left[1 - \text{Bernoulli}(\alpha_{x,Q|_{\{Y=x\}}})\right] \\ &= \int_A \alpha_{x,y} f_{Q|_{\{Y=x\}}}(y)\, dy + \mathbb{1}_{\{x\in A\}} \int_R (1 - \alpha_{x,y}) f_{Q|_{\{Y=x\}}}(y)\, dy \end{aligned} \tag{8.7}$$

©Nic Freeman, University of Sheffield, 2024.

where

$$\alpha_{x,y} = \min\left(1, \frac{f_{Y|_{\{Q=y\}}}(x) f_Y(y)}{f_{Q|_{\{Y=x\}}}(y) f_Y(x)}\right) \tag{8.8}$$

is such that $\alpha_{y_m,\tilde{y}}$ is precisely (8.1). The MH algorithm will have stationary distribution Y if and only if for all $A \subseteq R$,

$$\mathbb{P}[Y \in A] = \int_R p(x, A) f_Y(x)\, dy \tag{8.9}$$

The rest of the argument will concentrate on proving that (8.7) and (8.8) imply that (8.9) holds.

The choice of α in (8.8) is key. Our next goal is to show that

$$\alpha_{x,y} f_{Q|_{\{Y=x\}}}(y) f_Y(x) = \alpha_{y,x} f_{Y|_{\{Q=y\}}}(x) f_Y(y), \tag{8.10}$$

which by Lemma 1.6.1 is equivalent to

$$\alpha_{x,y} f_{Y,Q}(x, y) = \alpha_{y,x} f_{Y,Q}(y, x). \tag{8.11}$$

Using (8.8), this equation can be checked by considering two cases:

- if $f_{Y,Q}(x, y) \leq f_{Y,Q(y,x)}$ then $\alpha_{x,y} = 1$ and $\alpha_{y,x} = \frac{f_{Y,Q}(y,x)}{f_{Y,Q}(x,y)}$;
- if $f_{Y,Q}(x, y) \geq f_{Y,Q(y,x)}$ then $\alpha_{x,y} = \frac{f_{Y,Q}(x,y)}{f_{Y,Q}(y,x)}$ and $\alpha_{y,x} = 1$.

In both cases, (8.11) holds.

Remark 8.2.5 Recall the heuristic interpretation of probability density functions: $f_P(p)$ represents how likely P is to be close to p. From the MH algorithm, this means that the left hand side of (8.10) represents the likelihood of $y_{m+1} \approx x$ given that $y_m \approx y$, where y is sampled from Y, The right hand side of (8.10) represents the same concept but *with time run in reverse*, that is the likelihood of $y_{m+1} \approx y$ given that $y_m \approx x$, where x is sampled from Y. The choice of $\alpha_{x,y}$ in (8.8) ensures that these quantities are equal.

Equation (8.10) is closely related to *detailed balance* equations, which you have seen in earlier courses for discrete space chains. Loosely, (8.10) gives detailed balance equations for the chain conditional on the event that a proposal is accepted. The quantity $\alpha_{x,y}$ controls how likely a proposal for the jump $x \mapsto y$ is to be accepted, or equivalently how likely the chain is to stand still rather than move to y. Because $\alpha_{x,y}$ depends on y, this also controls how likely *all* of the various possible moves are, which in turn controls the stationary distribution.

We will now show that (8.9) holds. We have

$$\begin{aligned}
\int_R p(x, A) f_Y(x)\, dy &= \int_R \int_A \alpha_{x,y} f_{Q|_{\{Y=x\}}}(y) f_Y(x)\, dy\, dx + \int_R \int_R \mathbb{1}_{\{x \in A\}} (1 - \alpha_{x,y}) f_{Q|_{\{Y=x\}}} f_Y(x)\, dy\, dx \\
&= \int_R \int_A \alpha_{x,y} f_{Y,Q}(x, y)\, dy\, dx + \int_A \int_R (1 - \alpha_{x,y}) f_{Y,Q}(x, y)\, dy\, dx \\
&= \int_R \int_A \alpha_{y,x} f_{Y,Q}(y, x)\, dy\, dx + \int_A \int_R f_{Y,Q}(x, y)\, dy\, dx - \int_A \int_R \alpha_{x,y} f_{Y,Q}(x, y)\, dy\, dx \\
&= \int_A \int_R \alpha_{y,x} f_{Y,Q}(y, x)\, dx\, dy + \mathbb{P}[Y \in A, Q \in R] - \int_A \int_R \alpha_{x,y} f_{Y,Q}(x, y)\, dy\, dx \\
&= \mathbb{P}[Y \in A]
\end{aligned}$$

©Nic Freeman, University of Sheffield, 2024.

In the first line of the above we use (8.7) to expand $p(x, A)$. To obtain the second line we use Lemma 1.6.1. To obtain the third line we use (8.11) for the first term, and for other terms we simply split the integral into two. In the fourth line we exchange the order of integration in the first term, and note that the second term can be expressed as a probability. The final line follows because the first and third terms cancel (re-label x and y as each other in the first term to obtain the third) and because $\mathbb{P}[Q \in R] = 1$. We thus obtain (8.9), as required.

Remark 8.2.6 It is possible to weaken the conditions on (Y, Q) and allow cases where the range of $Q|_{\{Y=y\}}$ is a subset of the range of Y. In this case it becomes necessary that the random sequence (y_n) defined by the algorithm satisfies the condition $\mathbb{P}[\exists n \in \mathbb{N}, y_n \in A] = 1$ whenever $\mathbb{P}[Y \in A] > 0$, regardless of the starting point of the chain (y_n). This condition is known as *Harris recurrence*.

The same algorithm can also produce samples from discrete distributions. In this case we must replace the p.d.f f_Y by the p.m.f. p_Y, and similarly for the conditional parts in (8.1), but otherwise we proceed exactly as before. We have focused on continuous prior and posterior distributions, with the consequence that we won't need the discrete case of Metropolis-Hastings within this course.

©Nic Freeman, University of Sheffield, 2024.

8.3 Markov chain Monte Carlo

We will assume that the parameter space Π is a subset of $\mathbb{R}^d$. Recall from Theorems 2.4.1 and 3.1.2 that Π is also the range of both the prior Θ and the posterior $\Theta|_{\{X=x\}}$. As usual, we will assume that the prior distribution Θ is a continuous distribution with p.d.f. f_Θ.

We want to use the Metropolis-Hasting algorithm from Section 8.2 for the purposes of Bayesian inference. Our target is the posterior distribution $Y = \Theta|_{\{X=x\}}$. Let us think about what we capabilities we need, in order to do this.

1. We first need a choice of joint distribution (Y, Q), and the ability to take samples $\tilde{y}$ from the proposal distribution $Q|_{\{Y=y\}}$, for any value of y.

Random walk case:	**General case:**
Choose $Q = \Theta|_{\{X=x\}} + Z$, where Z satisfies $f_Z(z) = f_Z(-z)$, as in Example 8.2.4. Then $\widetilde{y} = y + Z$ has the distribution $Q|_{\{\Theta|_{\{X=x\}}=y\}}$. This only works if the parameter θ takes values in $\mathbb{R}$ (or more generally, $\mathbb{R}^d$).	If we can't use the symmetric case because e.g. θ takes values in some bounded or half-infinite interval, then there is a more difficult choice to make here.

2. The second requirement is that we can calculate the value of α in (8.1)/(8.4).

Random walk case:	**General case:**
We can calculate the p.d.f. $f_{\Theta|_{\{X=x\}}}$ using Theorems 2.4.1 and 3.1.2. In the symmetric case this is all that we need.	If we can't use the symmetric case, then we also need to evaluate $f_{Y|_{\{Q=q\}}}$ and $f_{Q|_{\{Y=y\}}}$, from our chosen joint distribution (Y, Q). Our choice of (Y, Q) (in the step above) will usually result in us being able to write down the joint p.d.f. $f_{Y,Q}$, from which we can use Lemma 1.6.1 to calculate $f_{Y|_{\{Q=q\}}}$ and $f_{Q|_{\{Y=y\}}}$.

3. The last thing that we require is that the proposals are accepted fairly often, in step 3 of the MH algorithm. As a rough guide, an acceptance rate of $15 - 50\%$ is generally viewed as sufficient.

   This happens naturally in many cases. When it does not, some additional techniques are required, some of which appear in the MSc material next semester.

We'll focus on the symmetric case, for the remainder of Section 8.3, because the choice of Q and associated complications can be quite difficult in the general case. We'll write down a full description of the MCMC algorithm for Bayesian inference, for the random walk case, in Section 8.3.1

©Nic Freeman, University of Sheffield, 2024.

8.3.1 MCMC algorithm for the random walk case

We start with a (discrete or continuous) Bayesian model (X, Θ), where the parameter space is $\Pi = \mathbb{R}^d$. We want to obtain samples of $\Theta|_{\{X=x\}}$, and we know the p.d.f. $f_{\Theta|_{\{X=x\}}}$ from Theorems 2.4.1/3.1.2.

We must first choose an initial point y_0. The value doesn't matter too much, anywhere 'in the middle' of Π or near the mass of Θ will do. We must also choose a continuous distribution for Z satisfying $f_Z(z) = f_Z(-z)$ for all $z \in \mathbb{R}$. A common choice is $Z \sim N(0, \sigma^2)$, as in Example 8.2.4.

Then, given y_m, we define y_{m+1} as follows.

1. Sample z from Z and set $\tilde{y} = y_m + z$.
2. Calculate $\alpha = \min\left(1, \frac{f_{\Theta|_{\{X=x\}}}(\tilde{y})}{f_{\Theta|_{\{X=x\}}}(y_m)}\right)$.
3. Then, set $y_{m+1} = \begin{cases} \tilde{y} & \text{with probability } \alpha, \\ y_m & \text{with probability } 1-\alpha. \end{cases}$

We repeat steps 1-3 until m is large enough that values taken by the y_m are no longer affected by the choice of y_0. This often has to be judged by eye, from a plot of the sequence (y_m).

Repeating the whole procedure obtains multiple samples of $\Theta|_{\{X=x\}}$, which we can plot in a histogram to get an approximation of the distribution. This is exactly what we already did in Example 8.2.3.

©Nic Freeman, University of Sheffield, 2024.

8.4 Gibbs sampling

Recall that our parameter θ may really be a vector of parameters, as in Section 4.5 where we considered the model $M_{(\mu,\sigma)} \sim \mathrm{N}(\mu, \sigma^2)$ with $\theta = (\mu, \sigma) \in \mathbb{R} \times (0, \infty)$. With the numerical techniques introduced in Section 8.3.1 it is possible to handle very complex models, which might have many parameters $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$.

For $d = 1$ or $d = 2$ the MH algorithm is effective. For much large d, what tends to happen is that it takes a very long time for the sequence (y_m) generated by the MH algorithm to explore the parameter space $\Pi \subseteq \mathbb{R}^d$, which means that it takes longer to converge to (the distribution of) $\Theta|_{\{X=x\}}$. This happens simply because there is *a lot* of space to explore inside $\mathbb{R}^d$ when d is large; this phenomenon is often known as the curse of dimensionality.

Example 8.4.1 Imagine you have parked your car inside a multi-story car park and then forgotten where you've parked it. If you know which level your car is on then you will only have to search on that level, and you will find your car in minutes. If you don't know which level, it will take you *much* longer. The first case is exploring $d = 2$, the second is $d = 3$. Actually, to make this example properly match the difference between $d = 2$ and $d = 3$, the car park should have the same number of floors as there are parking spaces along the side-length of each floor! The problem only gets worse in $d \geq 3$.

One strategy for working around this problem is to update the parameters $(\theta_1, \ldots, \theta_d)$ one at a time. That is, we would first change θ_1 while keeping $(\theta_2, \ldots, \theta_d)$ fixed, then we would change θ_2 while keeping $(\theta_1, \theta_3, \ldots, \theta_d)$ fixed, and so on. After updating θ_d we would then return to θ_1. It is helpful to introduce some notation for this: we write $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d)$ for the vector θ with the θ_i term removed. We use the same notation for the random vector Θ e.g. Θ_{-i}.

Remark 8.4.2 In reality, instead of updating the θ_i one-by-one, it is common to update the parameters in small batches. For example we might update $(\theta_1, \ldots, \theta_4)$ in one step, then $(\theta_5, \ldots, \theta_8)$ in the next step, and so on. It is helpful to put related parameters, with values that might strongly influence each other, within the same batch.

When we update the parameters in turn, a common choice of proposal distribution is to set $Q \sim \Theta_i|_{\{\Theta_{-i}=\theta_{-i},\, X=x\}}$ in the update for θ_i, where θ_{-i} are the values obtained from the previous update. This choice of proposal has the effect that, from (8.1), we end up with $\alpha = 1$ and all proposals are then accepted. When using proposals of this form, the MH algorithm is usually known as the *Gibbs sampler*.

Definition 8.4.3 The distributions of $\Theta_i|_{\{\Theta_{-i}=\theta_{-i},\, X=x\}}$, for $i = 1, \ldots, d$, are known as the *full conditional distributions*. From Lemma 1.6.1 the i^{th} full conditional has p.d.f. given by

$$\begin{aligned} f_{\Theta_i|_{\{\Theta_{-i}=\theta_{-i},\, X=x\}}}(\theta_i) &= \frac{f_{\Theta|_{\{X=x\}}}(\theta)}{\int_{\mathbb{R}^{d-1}} f_{\Theta|_{\{X=x\}}}(\theta_1, \ldots, \theta_{i-1}, \theta_i, \theta_{i+1}, \ldots, \theta_d)\, d\theta_{-i}} \\ &\propto f_{\Theta|_{\{X=x\}}}(\theta) \end{aligned} \tag{8.12}$$

We can calculate $f_{\Theta|_{\{X=x\}}}$ from Theorems 2.4.1/3.1.2, which provides a strategy for calculating (8.12) analytically. Note that $\propto$ in (8.12) treats θ_{-i} and x as constants, and the only variable is θ_i.

©Nic Freeman, University of Sheffield, 2024.

Remark 8.4.4 The notation $\Theta_i|_{\{\Theta_{-i}=\theta_{-i},\, X=x\}}$ for the full conditionals is a bit unwieldy. In Bayesian shorthand we would write simply $\theta_i|\theta_{-i}, x$ which is much neater.

The Gibbs sampler that results from these strategies is as follows.

1. Choose an initial point $y_0 = (\theta_1^{(0)}, \ldots, \theta_d^{(0)}) \in \Pi$.
2. For each $i = 1, \ldots, d$, sample $\tilde{y}$ from $\Theta_i|_{\{\Theta_{-i}=\theta_{-i}^{(m)},\, X=x\}}$ and set
$$y_{m+1} = (\theta_1^{(m)}, \ldots, \theta_{i-1}^{(m)}, \tilde{y}, \theta_{i+1}^{(m)}, \ldots, \theta_d^{(m)}).$$
Note that we increment the value of m each time that we increment i. When reach $i = d$, return to $i = 1$ and repeat.
Repeat this step until m is large enough that values taken by the y_m are no longer affected by the choice of y_0.
3. The final value of y_m is now a sample of $\Theta|_{\{X=x\}}$.

Note that we need to take samples from the full conditionals $\Theta_i|_{\{\Theta_{-i}=\theta_{-i}^{(m)},\, X=x\}}$ in step 2. This isn't always possible, and the Gibbs sampler is only helpful if we can do that. It is often used in cases where the full conditionals turn out to be named distributions, or nearly one as in Example 8.4.5 below.

For the MH algorithm we had Theorem 8.2.2 to tell us that y_m was (approximately) a sample of $\Theta|_{\{X=x\}}$, justified by the discussion in Section 8.2.3. It is possible to make similar arguments for the Gibbs algorithm above, but we won't include them in our course.

If the full conditionals can't be easily sampled from, then one strategy is to use the MH algorithm (run inside of step 2 above) to obtain samples of $\Theta_{-i}|_{\{X=x\}}$. This technique is known as *Metropolis-within-Gibbs*. In practice, once the parameters are divided up into batches, as described in Remark 8.4.2, some batches may be amenable to Gibbs sampling, whilst others may require Metropolis-within-Gibbs. The details will depend on the model. Trial and error is often required to find the best combination of techniques. We won't try to write down algorithms of that complexity within these notes, but you should hopefully end the course with an understanding of how (and why) each piece of an algorithm like that would work.

Example 8.4.5 This example comes from Sections 1.1.1/7.5.3/8.6.2 of the book 'Bayesian Approach to Intrepreting Archaeological Data' by Buck et al (1996). The data comes from radiocarbon dating, and is a vector $(x_1, x_2, \ldots, x_n)$ of estimated ages obtained (via carbon dating) from n different objects. We write θ_i for the true age of object i, which is unknown. Our model for the age of each object is $x_i \sim \mathrm{N}(\theta_i, v_i)$ and we assume that the estimation errors are independent, for each i. For simplicity we will assume that the v_i are known parameters, so our model family has n parameters $\theta = (\theta_1, \ldots, \theta_n)$. We thus have the model family
$$M_\theta = \mathrm{N}(\theta_1, v_1) \otimes \ldots \otimes \mathrm{N}(\theta_n, v_n).$$

From the historical context of the objects, it is known that $\theta_1 < \theta_2 < \ldots < \theta_n$, so we condition our model M_θ on this event. We can use Exercise 1.8 to do this conditioning, resulting in a new model family M'_θ given by
$$f_{M'_\theta}(x) = \begin{cases} \frac{f_{M_\theta}(x)}{\mathbb{P}[\mathrm{N}(\theta_1,v_1)<\mathrm{N}(\theta_2,v_2)<\ldots<\mathrm{N}(\theta_n,v_n)]} & \text{for } \theta_1 < \theta_2 < \ldots < \theta_n \\ 0 & \text{otherwise} \end{cases}$$

©Nic Freeman, University of Sheffield, 2024.

$$\propto \begin{cases} \prod_{i=1}^n f_{\mathrm{N}(\theta_i, v_i)}(x_i) & \text{for } \theta_1 < \theta_2 < \ldots < \theta_n \\ 0 & \text{otherwise} \end{cases}$$
$$\propto \begin{cases} \exp\left(-\frac{1}{2}\sum_{i=1}^n \frac{(\theta_i - x_i)^2}{v_i}\right) & \text{for } \theta_1 < \theta_2 < \ldots < \theta_n \\ 0 & \text{otherwise.} \end{cases}$$

We use the Bayesian model (X, Θ) with model family M'_θ and the improper prior

$$f_\Theta(\theta) = \begin{cases} 1 & \text{for } 0 < \theta_1 < \theta_2 < \ldots < \theta_n, \\ 0 & \text{otherwise.} \end{cases}$$

By Theorem 3.1.2 we obtain that the posterior distribution has p.d.f.

$$f_{\Theta|_{\{X=x\}}}(\theta) \propto \begin{cases} \exp\left(-\frac{1}{2}\sum_{i=1}^n \frac{(\theta_i - x_i)^2}{v_i}\right) & \text{for } 0 < \theta_1 < \theta_2 < \ldots < \theta_n, \\ 0 & \text{otherwise.} \end{cases} \tag{8.13}$$

This is the same density as M'_θ, except now we treat θ rather than x as the variable. The density is symmetric in x and θ so we already know this distribution. It is the distribution of $\theta \sim \mathrm{N}(x_1, v_1) \otimes \ldots \otimes \mathrm{N}(x_n, v_n)$ conditioned on the event $0 < \theta_1 < \theta_2 < \ldots < \theta_n$. One way to simulate samples of this distribution is via rejection sampling: simulate $\theta \sim \mathrm{N}(x_1, v_1) \otimes \ldots \otimes \mathrm{N}(x_n, v_n)$ and reject the sample θ until it satisfies $0 < \theta_1 < \theta_2 < \ldots < \theta_n$. The trouble is that unless n is small, we will mostly end up rejecting the samples because the condition we have imposed is an unlikely one.

From (8.13) and (8.12) we have full conditionals given by

$$f_{\Theta_i|_{\{\Theta_{-i}=\theta_{-i},\, X=x\}}}(\theta_i) \propto \begin{cases} \exp\left(-\frac{1}{2}\sum_{j=1}^n \frac{(\theta_j - x_i)^2}{v_i}\right) & \text{for } \theta_i \in (\theta_{i-1}, \theta_{i+1}), \\ 0 & \text{otherwise} \end{cases}$$
$$\propto \begin{cases} \exp\left(-\frac{1}{2}\frac{(\theta_i - x_i)^2}{v_i}\right) & \text{for } \theta_i \in (\theta_{i-1}, \theta_{i+1}), \\ 0 & \text{otherwise} \end{cases}$$

(where we set $\theta_0 = 0$ and $\theta_{n+1} = \infty$ to make convenient notation). Note that θ_i is the only variable here, and the second line follows because θ_{-i} and x are treated as constants. We recognize this full conditional distribution as that of $\theta_i \sim N(x_i, v_i)$ conditioned on the event $\theta_i \in (\theta_{i-1}, \theta_{i+1})$. These full conditionals are much easier to sample from: we use rejection sampling, sample $\theta_i \sim \mathrm{N}(x_i, v_i)$ and reject until we obtain a sample for which $\theta_i \in (\theta_{i-1}, \theta_{i+1})$. Hence, in this situation we have all the necessary ingredients to use a Gibbs sampler.

©Nic Freeman, University of Sheffield, 2024.

8.5 Exercises on Chapter 8

The exercises for this chapter, as well as the project that will comprise 15% of the MAS364 final marks, will appear (roughly) here within the online notes. The project will have at least some parts that do not depend on Chapter 8, so it may appear earlier on.

These notes have been written for first usage in Autumn 2024. I would rather wait and see how the course progresses before setting these parts, so these sections will only appear later on. We will discuss all this in lectures at the appropriate time.

©Nic Freeman, University of Sheffield, 2024.

Appendix A

Reference Sheets

The reference sheets displayed on the following pages will be provided in the exam. Full size copies will also be given out alongside these lecture notes, at the start of the course.

SOME DISCRETE DISTRIBUTIONS

Name	**Parameters**	**Genesis / Usage**	$p(x) = \mathbb{P}[X = x]$ **and non-zero range**	$\mathbb{E}[X]$	$\mathrm{Var}(X)$	**Comments**
Uniform (discrete)	$k \in \mathbb{N}$	Set of k equally likely outcomes.	$p(x) = 1/k$ $x = 1, ..., k$	$\frac{k+1}{2}$	$\frac{k^2-1}{12}$	Fair dice roll with $k = 6$.
Bernoulli trial	$\theta \in [0, 1]$	Experiment with two outcomes; typically, success = 1, fail = 0.	$p(x) = \theta^x(1-\theta)^{1-x}$ $x = 0, 1$	θ	$\theta(1-\theta)$	
Binomial	$n \in \mathbb{N}$ $\theta \in [0, 1]$	Number of successes in n i.i.d. Bernoulli trials.	$p(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ $x = 0, 1, 2, ..., n$	$n\theta$	$n\theta(1-\theta)$	Often written $\mathrm{Bin}(n, \theta)$. $\mathrm{Bin}(1, \theta) \sim \mathrm{Bernoulli}(\theta)$
Geometric	$\theta \in (0, 1]$	Number of failed i.i.d. Bernoulli trials before the first success.	$p(x) = \theta(1-\theta)^x$ $x = 0, 1, 2, \ldots$	$\frac{\theta}{1-\theta}$	$\frac{\theta^2}{(1-\theta)^2}$	Alternative parametrisations: swap θ and $1-\theta$, or $X' = X + 1$ to include the final trial.
Negative Binomial	$k \in \mathbb{N}$ $\theta \in (0, 1]$	Number of failed i.i.d. Bernoulli trials before the k^{th} success.	$p(x) = \binom{x+k-1}{x}\theta^k(1-\theta)^x$ $x = 0, 1, 2, \ldots$	$\frac{k(1-\theta)}{\theta}$	$\frac{k(1-\theta)}{\theta^2}$	Many alternative parametrisations. $\mathrm{NegBin}(1, \theta) \sim \mathrm{Geometric}(\theta)$.
Hypergeometric	$N \in \mathbb{N}$ $k \in \{0, \ldots, N\}$ $n \in \{0, \ldots, n\}$	Number of special objects in a random sample of n objects, from a population of N objects with k special objects.	$p(x) = \binom{k}{x}\binom{N-k}{n-x}/\binom{N}{n}$ $x = 0, ..., n$	$\frac{nk}{N}$	$n\frac{N-n}{N-1}\frac{k}{N}\times$ $(1-\frac{k}{N})$	
Poisson	$\lambda \in (0, \infty)$	Counting events occurring uniformly at random within space or time.	$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ $x = 0, 1, 2, ...$	λ	λ	

©Nic Freeman, University of Sheffield, 2024.

SOME CONTINUOUS DISTRIBUTIONS

Name	**Parameters**	**Genesis / Usage**	$f(x) =$ **p.d.f. and non-zero range**	$\mathbb{E}[X]$	$\mathrm{Var}(X)$	**Comments**
Uniform (continuous)	$\alpha, \beta \in \mathbb{R}$ with $\alpha < \beta$	The uniform distribution for a continuous interval.	$f(x) = \frac{1}{\beta-\alpha}$ $x \in (\alpha, \beta)$	$\frac{\alpha+\beta}{2}$	$\frac{(\beta-\alpha)^2}{12}$	
Normal	$\mu \in \mathbb{R}$ $\sigma \in (0,\infty)$	Empirically and theoretically (via CLT) a good model in many situations.	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ $x \in \mathbb{R}$	μ	σ^2	Often written $\mathrm{N}(\mu, \sigma^2)$. Alternative parameter: $\tau = \frac{1}{\sigma^2}$. $a\,\mathrm{N}(\mu, \sigma^2) + b \sim N(a\mu + b, a^2\sigma^2)$
Exponential	$\lambda \in (0,\infty)$	Inter-arrival times of random events.	$f(x) = \lambda e^{-\lambda x}$ $x \in (0,\infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	Often written $\mathrm{Exp}(\lambda)$. Alternative parameter: $\theta = \frac{1}{\lambda}$.
Gamma	$\alpha \in (0,\infty)$ $\beta \in (0,\infty)$	Lifetimes of ageing items, multi-inter-arrival times.	$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ $x \in (0,\infty)$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	Often written $\Gamma(\alpha, \beta)$. Alternative parameter: $\theta = \frac{1}{\beta}$. $\mathrm{Gamma}(1, \lambda) \sim \mathrm{Exp}(\lambda)$
Beta	$\alpha \in (0,\infty)$ $\beta \in (0,\infty)$	Quantities constrained to be within intervals.	$f(x) = \frac{1}{B(a,b)} x^{\alpha-1}(1-x)^{\beta-1}$ $x \in [0,1]$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$	$\mathrm{Beta}(1,1) \sim \mathrm{Uniform}(0,1)$
Cauchy	$a \in \mathbb{R}$ $b \in (0,\infty)$	Heavy tailed, pathological examples.	$f(x) = \frac{1}{\pi b} \frac{b^2}{(x-a)^2+b^2}$ $x \in \mathbb{R}$	undefined	undefined	
Pareto	$\alpha \in (0,\infty)$ $\beta \in (0,\infty)$	Heavy tailed quantities.	$f(x) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}}$ $x \in (\beta,\infty)$	$\frac{\alpha\beta}{\alpha-1}$ if $\alpha > 1$	$\frac{\alpha^2\beta}{(\alpha-1)^2(\alpha-2)}$ if $\alpha > 2$	Sometimes written $\mathrm{Pareto}(\beta, \alpha)$. $\log\left(\frac{\mathrm{Pareto}(\alpha,\beta)}{\beta}\right) \sim \mathrm{Exp}(\alpha)$
Weibull	$\lambda \in (0,\infty)$ $k \in (0,\infty)$	Lifetimes, extreme values.	$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$ $x \in (0,\infty)$	$\lambda\Gamma(1+1/k)$	$\lambda^2[\Gamma(1+2/k) + \Gamma(1+1/k)^2]$	$\left(\frac{\mathrm{Weibull}(\lambda,k)}{\lambda}\right)^k \sim \mathrm{Exp}(1)$
Log-Normal	$\mu \in \mathbb{R}$ $\sigma \in (0,\infty)$	Quantities related to exponential growth.	$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log x-\mu)^2}{\sqrt{2}\sigma}\right)$ $x \in (0,\infty)$	$e^{\mu+\frac{1}{2}\sigma^2}$	$(e^{\sigma^2}-1) \times e^{2\mu+\sigma^2}$	Often written $\mathrm{LogN}(\mu, \sigma^2)$. $\log(\mathrm{LogN}(\mu, \sigma^2)) \sim N(\mu, \sigma^2)$
Chi-squared	$n \in \mathbb{N}$	Statistical testing.	$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}$ $x \in (0,\infty)$	n	$2n$	Often written χ^2_n. $X^2_n \sim \mathrm{Gamma}(n/2, 1/2)$ $X_i \sim N(0,1)$ i.i.d. $\Rightarrow \sum_1^n X_i^2 \sim \chi^2_n$
Student t	$n \in \mathbb{N}$	Statistical testing.	$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1+\frac{x^2}{n})^{-\frac{n+1}{2}}$ $x \in \mathbb{R}$	0 if $n > 1$	$\frac{n}{n-2}$ if $n > 2$	Often written t_n. Can allow $n \in (0,\infty)$. $t_1 \equiv \mathrm{Cauchy}(0,1)$
Inverse Gamma	$\alpha \in (0,\infty)$ $\beta \in (0,\infty)$	Quantities related to the Gamma distribution.	$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp(-\beta/x)$ $x \in (0,\infty)$	$\frac{\beta}{\alpha-1}$ if $\alpha > 1$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ if $\alpha > 2$	Often written $\mathrm{IGamma}(\alpha, \beta)$. $\mathrm{IGamma}(\alpha, \beta) \sim \frac{1}{\mathrm{Gamma}(\alpha,\beta)}$

SOME CONJUGATE PAIRS

Model family	**Prior family**	**Data**	**Posterior parameters**
$\mathrm{Bernoulli}(\theta)^{\otimes n}$	$\theta \sim \mathrm{Beta}(a,b)$	$x \in \{0,1\}^n$	$a^* = a + \sum_1^n x_i$ $b^* = b + n - \sum_1^n x_i$
$\mathrm{Bin}(m_1,\theta) \otimes \ldots \otimes \mathrm{Bin}(m_n,\theta)$ with $m_1, \ldots m_n \in \mathbb{N}$ fixed.	$\theta \sim \mathrm{Beta}(a,b)$	$x \in \{0,1,\ldots,\}^n$ where $x_i \in \{0,\ldots,m_i\}$	$a^* = a + \sum_1^n x_i$ $b^* = b + \sum_1^n m_i - \sum_1^n x_i$
$\mathrm{Geometric}(\theta)^{\otimes n}$	$\theta \sim \mathrm{Beta}(a,b)$	$x \in \{0,1,\ldots,\}^n$	$a^* = a + n$ $b^* = b + \sum_1^n x_i$
$\mathrm{Poisson}(\theta)^{\otimes n}$	$\theta \sim \mathrm{Gamma}(a,b)$	$x \in \{0,1,\ldots,\}^n$	$a^* = a + \sum_1^n x_i$ $b^* = b + n$
$\mathrm{Exp}(\lambda)^{\otimes n}$	$\lambda \sim \mathrm{Gamma}(a,b)$	$x \in (0,\infty)^n$	$a^* = a + n$ $b^* = b + \sum_1^n x_i$
$\mathrm{Weibull}(\theta,\beta)^{\otimes n}$ with $\beta \in (0,\infty)$ fixed.	$\theta \sim \mathrm{IGamma}(a,b)$	$x \in (0,\infty)^n$	$a^* = a + n$ $b^* = b + \sum_1^n x_i^\beta$
$\mathrm{N}(\theta,\sigma^2)^{\otimes n}$ with $\sigma \in (0,\infty)$ fixed.	$\theta \sim \mathrm{N}(u,s^2)$	$x \in \mathbb{R}^n$	$u^* = \left(\frac{1}{\sigma^2}\sum_1^n x_i + \frac{u}{s^2}\right) / \left(\frac{n}{\sigma^2} + \frac{1}{s^2}\right)$ $(s^*)^2 = 1 / \left(\frac{n}{\sigma^2} + \frac{1}{s^2}\right)$
$\mathrm{N}(\theta,\frac{1}{\tau})^{\otimes n}$ with $\tau \in (0,\infty)$ fixed.	$\theta \sim \mathrm{N}(u,\frac{1}{t})$	$x \in \mathbb{R}^n$	$u^* = (\tau\sum_1^n x_i + ut) / (\tau n + t)$ $\frac{1}{t^*} = 1 / (\tau n + t)$
$\mathrm{N}(\mu,\frac{1}{\tau})^{\otimes n}$ with $\mu \in \mathbb{R}$ fixed.	$\tau \sim \mathrm{Gamma}(a,b)$	$x \in \mathbb{R}^n$	$a^* = a + \frac{n}{2}$ $b^* = b + \frac{1}{2}\sum_1^n (x_i - \mu)^2$
$\mathrm{N}(\mu,\frac{1}{\tau})^{\otimes n}$	$(\mu,\tau) \sim \mathrm{NGamma}(m,p,a,b)$	$x \in \mathbb{R}^n$	$m^* = \frac{n\bar{x}+mp}{n+p}$ $p^* = n + p$ $a^* = a + \frac{n}{2}$ $b^* = b + \frac{n}{2}\left(s^2 + \frac{p}{n+p}(\bar{x}-m)^2\right)$ where $\bar{x} = \frac{1}{n}\sum_1^n x_i$ and $s^2 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2$

See the sheet on conditional probability for the Normal-Gamma distribution.
For all other distributions, see the reference sheets of discrete and continuous distributions.

©Nic Freeman, University of Sheffield, 2024.

CONDITIONAL PROBABILITY AND RELATED FORMULAE

We say that a random variable X is **discrete** if there exists a countable set $A \subseteq \mathbb{R}^d$ such that $\mathbb{P}[X \in A] = 1$. In this case the function $p_X(x) = \mathbb{P}[X = x]$, defined for $x \in \mathbb{R}^d$, is known as the **probability mass function** of X. The **range** of X is the set $R_X = \{x \in \mathbb{R}^d ; \mathbb{P}[X = x] > 0\}$.

We say that a random variable X is **continuous** if there exists a function $f_X : \mathbb{R}^d \to [0, \infty)$ such that $\mathbb{P}[X \in A] = \int_A f_X(x)\, dx$ for all $A \subseteq \mathbb{R}^d$. In this case f_X is known as the **probability density function** of X. The **range** of X is the set $R_X = \{x \in \mathbb{R}^d ; f_X(x) > 0\}$.

If X and Y are discrete, and $p_X \propto p_Y$, then $X \overset{\mathrm{d}}{=} Y$.
If X and Y are continuous, and $f_X \propto f_Y$, then $X \overset{\mathrm{d}}{=} Y$.

If X is a random variable and $\mathbb{P}[X \in A] > 0$ then the **conditional distribution** of $X|_{\{X\in A\}}$ satisfies $\mathbb{P}[X|_{\{X\in A\}} \in A] = 1$ and

$$\mathbb{P}[X|_{\{X\in A\}} \in B] = \frac{\mathbb{P}[X \in B]}{\mathbb{P}[X \in A]}$$

for all $B \subseteq A$.

If X and Y are random variables, with $A \subseteq R_X$, $B \subseteq R_Y$ and $\mathbb{P}[X \in A] > 0$, then

$$\mathbb{P}[Y|_{\{X\in A\}} \in B] = \frac{\mathbb{P}[X \in A, Y \in B]}{\mathbb{P}[X \in A]}.$$

If (Y, Z) and random variables and $\mathbb{P}[Y = y] = 0$ then it is sometimes possible to define the conditional distribution of $Z|_{\{Y=y\}}$ via taking the limit $\mathbb{P}\left[Z|_{\{|Y-y|\le\epsilon\}} \in A\right] \to \mathbb{P}[Z|_{\{Y=y\}} \in A]$ as $\epsilon \to 0$.

Let (Y, Z) be a pair of continuous random variables. If the conditional distribution of $Z|_{\{Y=y\}}$ exists then it is given by

$$f_{Z|_{\{Y=y\}}}(z) = \frac{f_{Y,Z}(y, z)}{f_Y(y)}.$$

For a discrete or continuous random variable X, the **likelihood function** of X is

$$L_X(x) = \begin{cases} \mathbb{P}[X = x] & \text{if } X \text{ is discrete,} \\ f_X(X) & \text{if } X \text{ is continuous.} \end{cases}$$

The general formula for **completing the square** as a function of $\theta \in \mathbb{R}$ is $A\theta^2 - 2\theta B + C = A\left(\theta - \frac{B}{A}\right)^2 + C - \frac{B^2}{A}$

The **sample-mean-variance** identity states $\sum_1^n (x_i - \mu)^2 = ns^2 + n(\bar{x} - \mu)^2$ where $\bar{x} = \frac{1}{n}\sum_1^n x_i$ and $s^2 = \frac{1}{n}\sum_1^n (x_i - \bar{x})^2$.

The **Beta and Gamma functions** are given by

$$\mathcal{B}(a, b) = \int_0^1 x^{a-1}(1 - x)^{b-1}\, dx, \qquad \Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\, dt.$$

They are related by $\mathcal{B}(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. For $n \in \mathbb{N}$, $(n - 1)! = \Gamma(n)$.

The **Normal-Gamma distribution** has p.d.f. given by

$$\begin{aligned} f_{\mathrm{NGamma}(m,p,a,b)}(\mu, \tau) &= f_{\mathrm{N}(m,\frac{1}{p\tau})}(\mu)\, f_{\mathrm{Gamma}(a,b)}(\tau) \\ &\propto \tau^{a-\frac{1}{2}} \exp\left(-\frac{p\tau}{2}(\mu - m)^2 - b\tau\right). \end{aligned}$$

for $\mu \in \mathbb{R}$ and $\tau > 0$, and zero otherwise. The parameters are $m \in \mathbb{R}$, $p \in (0, \infty)$, $a \in (0, \infty)$ and $b \in (0, \infty)$. If $(U, T) \sim \mathrm{NGamma}(m, p, a, b)$ then $T \sim \mathrm{Gamma}(a, b)$ and $U|_{\{T=\tau\}} \sim \mathrm{N}(m, \frac{1}{p\lambda})$.

BAYESIAN MODELS AND RELATED FORMULAE

The **Bayesian model** associated to the model family $(M_\theta)_{\theta\in\Pi}$ and prior p.d.f. $f_\Theta(\theta)$ is the random variable $(X, \Theta) \in \mathbb{R}^n \times \mathbb{R}^d$ with distribution given by

$$\mathbb{P}[X \in B, \Theta \in A] = \int_A \mathbb{P}[M_\theta \in B] f_\Theta(\theta)\, d\theta.$$

The model family satisfies $X|_{\{\Theta=\theta\}} \overset{\mathrm{d}}{=} M_\theta$.

The distribution of X is known as the **sampling distribution**, given by

$$\begin{aligned} \mathbb{P}[X = x] &= \int_{\mathbb{R}^d} \mathbb{P}[M_\theta = x] f_\Theta(\theta)\, d\theta \qquad \text{if } (M_\theta) \text{ is a discrete family,} \\ f_X(x) &= \int_{\mathbb{R}^d} f_{M_\theta}(x) f_\Theta(\theta)\, d\theta. \qquad \text{if } (M_\theta) \text{ is a continuous family.} \end{aligned} \tag{$\star$}$$

The distribution of $\Theta|_{\{X=x\}}$ is known as the **posterior distribution** given the data x. **Bayes rule** states that

$$f_{\Theta|_{\{X=x\}}}(\theta) = \frac{1}{Z} L_{M_\theta}(x) f_\Theta(\theta)$$

where L_{M_θ} is the likelihood function of M_θ; the p.d.f. in the absolutely continuous case and the p.m.f. in the discrete case. The normalizing constant Z is given by $Z = \int_\Pi L_{M_\theta}(x) f_\Theta(\theta)\, d\theta$, which is equal to $\mathbb{P}[X = x]$ in the discrete case and equal to $f_X(x)$ is the continuous case.

The **predictive distribution** is given by replacing f_Θ in $(\star)$ with $f_{\Theta|_{\{X=x\}}}$.

If θ is a real valued parameter and $X \sim M_\theta$, the **reference prior** Θ associated to the model family (M_θ) has density function given by

$$f_\Theta(\theta) \propto \mathbb{E}\left[\left(\left(\frac{d}{d\theta}\log(L_{M_\theta}(X))\right)^2\right]^{1/2} \propto \mathbb{E}\left[-\frac{d^2}{d\theta^2}\log(L_{M_\theta}(X))\right]^{1/2}.$$

Consider a Bayesian model with unknown parameter θ and data x. Let H_0 be the hypothesis that $\theta \in \Pi_0$, and H_1 be the hypothesis that $\theta \in \Pi_1$, where Π_0 and Π_1 partition the parameter space Π. The **prior and posterior odds ratios** of H_0 against H_1 are

$$\frac{\mathbb{P}[\Theta \in \Pi_0]}{\mathbb{P}[\Theta \in \Pi_1]} \qquad \text{and} \qquad \frac{\mathbb{P}[\Theta|_{\{X=x\}} \in \Pi_0]}{\mathbb{P}[\Theta|_{\{X=x\}} \in \Pi_1]}.$$

The **Bayes factor** is $B = \frac{\text{posterior odds}}{\text{prior odds}}$. The following table provides a rough guide to interpreting the Bayes factor.

Bayes factor	Interpretation: evidence in favour of H_0 over H_1
1 to 3.2	Indecisive / not worth more than a bare mention
3.2 to 10	Substantial
10 to 100	Strong
above 100	Decisive

A **high posterior density region** is a subset $\Pi_0 \subseteq \Pi$ that is chosen to minimize the size of Π_0 and maximize $\mathbb{P}[\Theta|_{\{X=x\}} \in \Pi_0]$.

If $\Theta|_{\{X=x\}}$ has a distribution with a single peak then it is common to choose an **equally tailed** HPD region of the form $\Pi_0 = [a, b]$ where

$$\mathbb{P}\left[\Theta|_{\{X=x\}} < a\right] = \mathbb{P}\left[\Theta|_{\{X=x\}} > b\right] = \frac{1 - p}{2}$$

and some value is picked for $p \in (0, 1)$.

If $Z \sim \mathrm{N}(0, 1)$ then $\mathbb{P}[Z \ge 1.645] \approx 0.95$, $\mathbb{P}[Z \ge 1.96] \approx 0.975$ and $\mathbb{P}[Z \ge 2.58] \approx 0.995$.

©Nic Freeman, University of Sheffield, 2024.

SOME USEFUL ALGORITHMS

The **Metropolis-Hastings** algorithm for simulating (approximate) samples from the distribution of Y is as follows. The key ingredient of the algorithm is a joint distribution (Y, Q), where $Q|_{\{Y=y\}}$ and $Y|_{\{Q=y\}}$ are both well defined for all $y \in R_Y$, both with the same range as Y.

Let y_0 be a point within R_Y. Then, given y_m we define y_{m+1} as follows.

1. Generate a *proposal point* $\tilde{y}$ from the distribution of $Q|_{\{Y=y_m\}}$.
2. Calculate the value of $\alpha = \min\left\{1, \frac{f_{Y|_{\{Q=\tilde{y}\}}}(y_m) f_Y(\tilde{y})}{f_{Q|_{\{Y=y_m\}}}(\tilde{y}) f_Y(y_m)}\right\}$.
3. Then, set $y_{m+1} = \begin{cases} \tilde{y} & \text{with probability } \alpha, \\ y_m & \text{with probability } 1-\alpha. \end{cases}$

For sufficiently large m, the distribution of y_m is approximately that of Y.

The distribution $Q|_{\{Y=y\}}$ is called the *proposal* distribution, based on its role in steps 1 and 2. The two cases in step 3 are usually referred to as *acceptance* (when $y_{m+1} = \tilde{y}$) and *rejection* (when $y_{m+1} = y_m$).

The **Metropolis** algorithm is the special case

$$f_{Q|_{\{Y=y\}}}(\tilde{y}) = f_{Y|_{\{Q=\tilde{y}\}}}(y), \tag{†}$$

in which case step 2 simplifies to $\alpha = \min\left\{1, \frac{f_Y(\tilde{y})}{f_Y(y_m)}\right\}$.

The **random walk Metropolis** algorithm is the choice $Q = Y + Z$, where Z is independent of Y and Q and satisfies $f_Z(z) = f_Z(-z)$ for all $z \in R_Z$. In this case

$$Q|_{\{Y=y\}} \overset{\mathrm{d}}{=} y + Z \qquad \text{and} \qquad Y|_{\{Q=\tilde{y}\}} \overset{\mathrm{d}}{=} \tilde{y} + Z,$$

which implies (†). A common choice is $Z \sim \mathrm{N}(0, \sigma^2)$.

The **random walk MCMC algorithm** is obtained by applying the random walk Metroplis algorithm to find the posterior distribution of a Bayesian model. The algorithm is as follows. We start with a (discrete or continuous) Bayesian model (X, Θ), where the parameter space is $\Pi = \mathbb{R}^d$. We want to obtain samples of $\Theta|_{\{X=x\}}$ and we know the p.d.f. $f_{\Theta|_{\{X=x\}}}$.

Choose an initial point $y_0 \in \Pi$. Choose a continuous distribution for Z satisfying $f_Z(z) = f_Z(-z)$ for all $z \in \mathbb{R}$. A common choice is $Z \sim N(0, \sigma^2)$.

Then, given y_m, we define y_{m+1} as follows.

1. Sample z from Z and set $\tilde{y} = y_m + z$.
2. Calculate $\alpha = \min\left(1, \frac{f_{\Theta|_{\{X=x\}}}(\tilde{y})}{f_{\Theta|_{\{X=x\}}}(y_m)}\right)$.
3. Then, set $y_{m+1} = \begin{cases} \tilde{y} & \text{with probability } \alpha, \\ y_m & \text{with probability } 1-\alpha. \end{cases}$

The **Gibbs sampler** for $\theta = (\theta_1, \ldots, \theta_d)$ is as follows. We first choose an initial point $y_0 = (\theta_1^{(0)}, \ldots, \theta_d^{(0)}) \in \Pi$. Then, for each $i = 1, \ldots, d$, sample $\tilde{y}$ from $\Theta_{-i}|_{\{X=x\}}$ and set

$$y_{m+1} = (\theta_1^{(m)}, \ldots, \theta_{i-1}^{(m)}, \tilde{y}, \theta_{i+1}^{(m)}, \ldots, \theta_d^{(m)}).$$

Note that we increment the value of m each time that we increment i. When reach $i = d$, return to $i = 1$ and repeat. For sufficiently large m, the distribution of y_m is approximately that of $\Theta|_{\{X=x\}}$.

The distributions of $\Theta_i|_{\{\Theta_{-i}=\theta_{-i}, X=x\}}$, for $i = 1, \ldots, d$, are known as the **full conditional distributions** of Θ. They satisfy

$$f_{\Theta_i|_{\{\Theta_{-i}=\theta_{-i}, X=x\}}}(\theta_i) \propto f_{\Theta|_{\{X=x\}}}(\theta)$$

Here $\propto$ treats θ_{-i} and x as constants, and the only variable is θ_i.

©Nic Freeman, University of Sheffield, 2024.

Appendix B

Advice for revision/exams

There are two different exam papers, one for MAS364 (sat in January) and one for MAS61006 (sat in the summer). The MAS61006 exam contains additional questions on the material taught in semester 2. For both exams the rubric reads:

> *Candidates should attempt ALL questions. The maximum marks for the various parts of the questions are indicated.*

Within these notes, material marked with a (⊘) is non-examinable for everyone. You do not need to study these parts during your revision.

- You will be asked to solve problems based on the material in these notes. There will be a broad range of difficulty amongst the questions. Some will be variations of questions in the assignments/notes, others will also try to test your ingenuity.
- Many of the important definitions and results appear on the (six page long!) reference sheets. You should practice using the reference sheets to help you solve problems.
- You will not be expected to reproduce long proofs from memory. Most proofs are marked as off-syllabus anyway, within these notes. You are expected to have followed the techniques within the proofs when they are present, and to be able to use these techniques in your own problem solving (e.g. Lemma 1.5.1).
- There are marks for attempting a suitable method, and for justifying mathematical steps, as well as for reaching a correct conclusion.

Revision activities

The most important activities:

1. Solve the assignment questions, and end-of-chapter exercises that are at ⋆ and ⋆⋆ difficulty levels. Check your solutions against the typed solutions.
2. Learn the key definitions, results, and examples.
3. Do the practice exam paper and mark your own solutions.

In all cases, you are welcome to come and discuss any questions/comments/typos. Please use office hours or email to arrange a convenient time.

©Nic Freeman, University of Sheffield, 2024.

Appendix C

Solutions to exercises

Chapter 1

1.1 (a) Samples of X will tend to be in one of three different locations: (1) sharply clustered around -7.5, (2) a broad cluster between approximately $[-4, 4]$ and (3) close to 10, but not greater than 10.

(b) We have

$$\int_{\mathbb{R}} f_{X_\theta}(x)\,dx = \int_1^\infty (\theta-1)x^{-\theta}\,dx = \left[(\theta-1)\frac{x^{1-\theta}}{1-\theta}\right]_{x=1}^{\infty} = 1$$

as required.

1.2 Neither.

To see that Z is not continuous: recall that for any continuous random variable Z' we have $\mathbb{P}[Z' = z] = 0$ for all z'. However, $\mathbb{P}[Z = 0] = \mathbb{P}[Y = 0] + \mathbb{P}[Y = 1, X = 0] = \frac{1}{2} + \frac{1}{2}(0) = \frac{1}{2}$.

To see that Z is not discrete: note that Z takes values across all of $\mathbb{R}$, coming from the case where $Y = 1$, which has probability $\frac{1}{2}$.

1.3 (a) U' has the conditional distribution of U given the event $\{U \in [a, b]\}$.

(b) We apply Lemma 1.4.1 with $A = [a, b]$. Part 1 gives that $\mathbb{P}[U' \in [a, b]] = 1$. Part 2 gives that for all $B \subseteq [a, b]$ we have

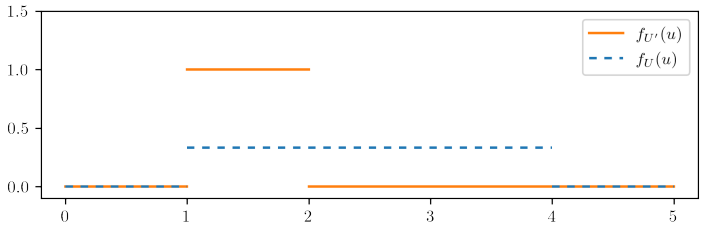
$$\mathbb{P}[U' \in B] = \frac{\mathbb{P}[U \in B]}{\mathbb{P}[U \in A]} = \frac{\int_B \frac{1}{c-a}\,dx}{\int_a^b \frac{1}{c-a}\,dx} = \frac{\int_B \frac{1}{c-a}\,dx}{\frac{b-a}{c-a}} = \int_B \frac{1}{b-a}\,dx.$$

Hence U' is a continuous random variable with p.d.f.

$$f_{U'}(u) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise.} \end{cases}$$

This is the continuous uniform distribution on $[a, b]$.

(c) We have:



1.4 Suppose that G has the Geometric(p) distribution, that is $\mathbb{P}[G = g] = p^{g-1}(1-p)$ for $g \in \{1, \ldots\}$, where $p \in [0, 1]$. Let $G' \overset{\mathrm{d}}{=} G|_{\{G \geq n\}}$, where $n \in \mathbb{N}$.

(a) We apply Lemma 1.4.1 with $A = \{n, n+1, \ldots\}$. From part 1 of the lemma we have $\mathbb{P}[G' \in \{n, n+1, \ldots, \}] = 1$. From part 2, for $g \in \{n, n+1, \ldots\}$ we have

$$\mathbb{P}[G' = g] = \frac{\mathbb{P}[G = g]}{\mathbb{P}[G \in A]} = \frac{p^g(1-p)}{\sum_{k=n}^\infty p^k(1-p)} = \frac{p^g(1-p)}{(1-p)\sum_{k=n}^\infty p^k} = \frac{p^g(1-p)}{(1-p)\frac{p^n}{1-p}} = p^{g-n}(1-p)$$

©Nic Freeman, University of Sheffield, 2024.

(b) The claim is correct. For $g \in \{n, n+1, \dots, \}$, the $-n$ term in p^{g-n} above corresponds to removing the factors p corresponding to success/failure of the first n trials, from what would otherwise have been p^g in the p.m.f. of G.

1.5 You should notice that as n gets larger it takes longer to obtain samples, and it quickly becomes impractical to do so as n grows large. This is because it becomes more likely that samples fall into $(-\infty, n)$ and are rejected, so it takes longer to find a sample that is accepted.

1.6 (a) We have

$$\sum_{y=1}^{\infty}\sum_{x=1}^{\infty} 2^{-xy}(1-2^{-y}) = \sum_{y=1}^{\infty}(1-2^{-y})\sum_{x=1}^{\infty}(2^{-y})^x = \sum_{y=1}^{\infty}(1-2^{-y})\frac{2^{-y}}{1-2^{-y}} = \sum_{y=1}^{\infty}2^{-y} = \frac{1/2}{1-1/2} = 1$$

as required. Here we use that the summations are geometric sums.

The random variables X and Y are not independent because (1.14) does not factorise into the form $g(x)h(y)$.

(b) (i) To find the marginal distribution of Y we sum over all possible values of x, giving

$$\mathbb{P}[Y=y] = \sum_{x=1}^{\infty} 2^{-xy}(1-2^{-y}) = (1-2^{-y})\sum_{x=1}^{\infty}(2^{-y})^x = (1-2^{-y})\frac{2^{-y}}{1-2^{-y}} = \left(\frac{1}{2}\right)^y$$

for $y \in \mathbb{N}$.

(ii) Using Lemma 1.5.1 we have

$$\mathbb{P}[X|_{\{Y=5\}} = x] = \frac{\mathbb{P}[X=x, Y=5]}{\mathbb{P}[Y=5]} = \frac{2^{-5x}(1-2^{-5})}{2^{-5}} = \left(1-\frac{1}{2^5}\right)\left(\frac{1}{2^5}\right)^{x-1}.$$

In the middle equality we use (1.14) and part (a).

(iii) Again using Lemma 1.5.1 we have

$$\mathbb{P}[Y|_{\{X\geq 5\}} = y] = \frac{\mathbb{P}[X\geq 5, Y=y]}{\mathbb{P}[X\geq 5]}. \tag{C.1}$$

We need to calculate the top and bottom of (C.1). For $y \in \mathbb{N}$,

$$\begin{aligned}\mathbb{P}[X\geq 5, Y=y] &= \sum_{x=5}^{\infty} 2^{-xy}(1-2^{-y}) \\ &= (1-2^{-y})\sum_{x=5}^{\infty}(2^{-y})^x = (1-2^{-y})\frac{(2^{-y})^5}{1-2^{-y}} = 2^{-5y}.\end{aligned}$$

Hence $\mathbb{P}[X\geq 5] = \sum_{y=1}^{\infty} 2^{-5y} = \sum_{y=1}^{\infty}(2^{-5})^y = \frac{2^{-5}}{1-2^{-5}}$. Putting these into (C.1) we obtain

$$\mathbb{P}[Y|_{\{X\geq 5\}} = y] = \frac{2^{-5y}(1-2^{-5})}{2^{-5}} = \left(1-\frac{1}{2^5}\right)\left(\frac{1}{2^5}\right)^{y-1}$$

for $y \in \mathbb{N}$.

The distributions found in (b) are all Geometric distributions. They have range $\{1, 2, \dots, \}$ rather than $\{0, 1, \dots, \}$ i.e. using the alternative parametrization mentioned on the reference sheet.

1.7 Let $X \in \mathrm{N}(0,1)$ and set $A = [0, \infty)$, as in Example 1.4.3. Let $Y' = |X|$. Show that $Y' \overset{\mathrm{d}}{=} X|_{\{X\in A\}}$.

Note that $\mathbb{P}[Y' \geq 0] = 1$. For $y > 0$ we can calculate,

$$\mathbb{P}[Y' \leq y] = \mathbb{P}[X \geq -y \text{ or } X \leq y] = \int_{-y}^{y} f_X(x)\,dx = \int_{-y}^{0} f_X(x)\,dx + \int_{0}^{y} f_X(x)\,dx = 2\int_{0}^{y} f_X(x).$$

The last equality follows by symmetry (or a $v = -y$ substitution) because $f_X(x) = f_X(-x)$. Differentiating, for $y > 0$ we have $f_{Y'}(y) = 2f_X(y)$. Hence Y' is a continuous random variable with p.d.f.

$$f_{Y'}(y) = \begin{cases} 2f_X(y) & \text{for } y > 0 \\ 0 & \text{otherwise.} \end{cases}$$

This matches the distribution of $X|_{\{X\geq 0\}}$ that we obtained in Example 1.4.3.

©Nic Freeman, University of Sheffield, 2024.

1.8 From part 2 of Lemma 1.4.1, for $B \subseteq A$ we have

$$\mathbb{P}[X|_{\{X\in A\}} \in B] = \frac{\mathbb{P}[X \in A \cap B]}{\mathbb{P}[X \in A]} = \frac{\mathbb{P}[X \in B]}{\mathbb{P}[X \in A]} = \frac{\int_B f_X(x)\,dx}{\mathbb{P}[X \in A]} = \int_B \frac{f_X(x)}{\mathbb{P}[X \in A]}\,dx.$$

By part 1 of Lemma 1.4.1 we have $\mathbb{P}[X|_{\{X\in A\}} \in B] = 1$. By Definition 1.1.1 X is a continuous random variable with p.d.f.

$$f_{X|_{\{X\in A\}}}(x) = \begin{cases} \frac{f_X(x)}{\mathbb{P}[X\in A]} & \text{if } x \in A \\ 0 & \text{otherwise.} \end{cases}$$

1.9 Let us write $\mathcal{L}_X(A) = \mathbb{P}[X \in A]$ for the law of X.

For the 'if' part, suppose that $X \overset{\mathrm{d}}{=} Y$. Take $A = \{x\}$ in Definition 1.2.1, where $x \in R$, then $\mathcal{L}_X(\{x\}) = \mathbb{P}[X \in \{x\}] = \mathbb{P}[X = x] = p_X(x)$, and similarly for Y. Since $\mathcal{L}_X = \mathcal{L}_Y$ we have $p_X(x) = p_Y(x)$.
For the 'only if' part, suppose that $p_X(x) = p_Y(x)$ for all $x \in R^d$. Note that for any $A \subseteq \mathbb{R}^d$ we have $\mathcal{L}_X(A) = \mathbb{P}[X \in A] = \sum_{x\in A} \mathbb{P}[X = a] = \sum_{x\in A} p_X(x)$, and similarly for Y. Hence $\mathcal{L}_X(A) = \mathcal{L}_Y(A)$.

1.10 Suppose that X takes values in $\mathbb{R}^n$ and Y takes values in $\mathbb{R}^d$. We apply Lemma 1.4.1, conditioning (X, Y) to be inside the set $A \times \mathbb{R}^d$. By part 2 of that lemma, for all $B \subseteq \mathbb{R}^d$ we have

$$\mathbb{P}[(X,Y) \in A \times B] = \frac{\mathbb{P}[(X,Y) \in A \times B]}{\mathbb{P}[(X,Y) \in A \times \mathbb{R}^d]} = \frac{\mathbb{P}[X \in A, Y \in B]}{\mathbb{P}[X \in A]} = \frac{\mathbb{P}[X \in A]\mathbb{P}[Y \in B]}{\mathbb{P}[X \in A]} = \mathbb{P}[Y \in B], \quad \text{(C.2)}$$

where we have used the fact that X and Y are independent. By part 1 of the lemma we have $\mathbb{P}[(X,Y)|_{\{X\in A\}} \in A \times \mathbb{R}^d] = 1$, which since $(X,Y)|_{\{X\in A\}} = (X|_{\{X\in A\}}, Y|_{\{X\in A\}})$ means that $\mathbb{P}[X|_{\{X\in A\}} \in A] = 1$. Hence, for all $B \subseteq \mathbb{R}^d$

$$\mathbb{P}[Y|_{\{X\in A\}} \in B] = \mathbb{P}[X|_{\{X\in A\}} \in A \text{ and } Y|_{\{X\in A\}} \in B] = \mathbb{P}[(X,Y)|_{\{X\in A\}} \in A \times B] = \mathbb{P}[Y \in B].$$

The last equality above uses (C.2). Thus $Y \overset{\mathrm{d}}{=} Y|_{\{X\in A\}}$.

Chapter 2

2.1 See `2_dist_sketching_solution.ipynb` or `2_dist_sketching_solution.Rmd`

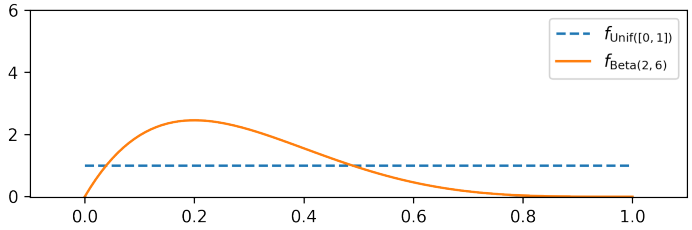
2.2 (a) $\mathbb{P}[M_p = n] = p(1-p)^n$ for $n \in \{1, 2, \dots, \}$.

(b) From Theorem 2.4.1 the posterior distribution is given by

$$\begin{aligned} f_{P|_{\{X=5\}}}(p) &= \frac{1}{\int_0^1 \mathbb{P}[\text{Geometric}(q) = 5] f_{\text{Uniform}([0,1])}(q)\,dq} \mathbb{P}[\text{Geometric}(p) = 5] f_{\text{Uniform}([0,1])}(p)\,dp \\ &= \frac{1}{\int_0^1 q(1-q)^5\,dq} p(1-p)^5 \\ &= \frac{1}{\mathcal{B}(2,6)} p(1-p)^5 \end{aligned}$$

for $p \in [0,1]$ and zero elsewhere, which we recognize as the p.d.f. of Beta$(2,6)$.

(c) We obtain



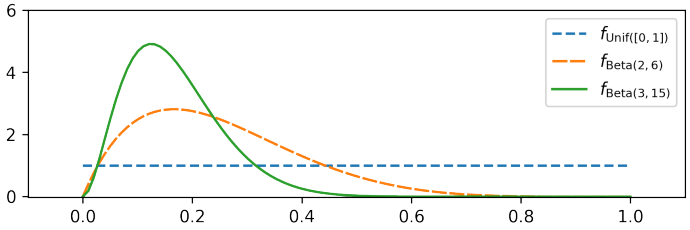
(d) From Theorem 2.4.1, now with the prior taken as $P \sim$ Beta$(2,6)$ and the new data point $x = 9$, the posterior distribution is given by

$$f_{P|_{\{X=9\}}}(p) = \frac{1}{\int_0^1 \mathbb{P}[\text{Geometric}(q) = 9] f_{\text{Beta}(2,6)}(q)\,dq} \mathbb{P}[\text{Geometric}(p) = 9] f_{\text{Beta}(2,6)}(p)\,dp$$

©Nic Freeman, University of Sheffield, 2024.

$$= \frac{\mathcal{B}(2,6)}{\mathcal{B}(2,6)} \frac{1}{\int_0^1 q(1-q)^5 q(1-q)^9 q\, dq} p(1-p)^5 p(1-p)^9$$
$$= \frac{1}{\mathcal{B}(3,15)} p^2(1-p)^{14}$$

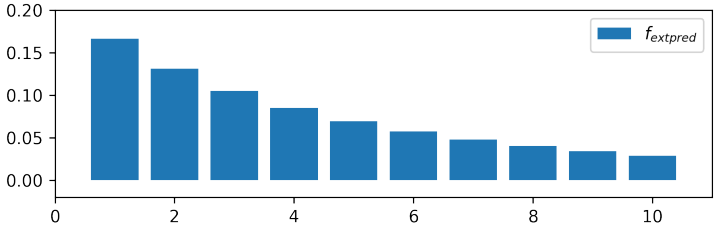
which we recognize as the p.d.f. of Beta$(3,15)$. Including this into our graph from (c),



(e) The p.m.f. of the predictive distribution is

$$\mathbb{P}[X' = x'] = \int_0^1 \mathbb{P}[\text{Geometric}(p) = x'] f_{\text{Beta}(3,15)}(p)\, dp$$

for $x' \in \{0, 1, \ldots\}$. We sketch this:



2.3 From Theorem 2.4.1 the posterior has p.d.f.

$$f_{\Lambda|\{X=5\}}(\lambda) = \frac{1}{\int_0^\infty \mathbb{P}[\text{Poisson}(l) = 7] f_{\text{Exp}(5)}(l)\, dl} \mathbb{P}[\text{Poisson}(\lambda) = 7] f_{\text{Exp}(5)}(\lambda)$$
$$= \frac{7!}{7!} \frac{1}{\int_0^\infty l^7 e^{-l} l e^{-5l}\, dl} \lambda^7 e^{-\lambda} \lambda e^{-5\lambda}$$
$$= \frac{1}{\int_0^\infty l^8 e^{-6l}\, dl} \lambda^8 e^{-6\lambda}$$
$$= \frac{6^9}{\Gamma(9)} \frac{1}{\int_0^\infty \frac{6^9}{\Gamma(9)} l^8 e^{-6l}\, dl} \lambda^8 e^{-6\lambda}$$
$$= \frac{6^9}{\Gamma(9)} \frac{1}{\int_0^\infty f_{\Gamma(6,9)(l)}\, dl} \lambda^8 e^{-6\lambda}$$
$$= \frac{6^9}{\Gamma(9)} \lambda^8 e^{-6\lambda}$$

for $\lambda > 0$ and zero otherwise. We recognize this as the p.d.f. of the Gamma$(9,6)$ distribution. The predictive p.m.f. is given by

$$\mathbb{P}[X' = x] = \int_0^\infty \mathbb{P}[\text{Poisson}(\lambda) = x] f_{\text{Gamma}(9,6)}(\lambda)\, d\lambda$$
$$= \int_0^\infty \frac{\lambda^x e^{-\lambda}}{x!} \frac{9^6}{\Gamma(9)} \lambda^8 e^{-6\lambda}\, d\lambda$$
$$= \frac{9^6}{8! x!} \int_0^\infty \lambda^{8+x} e^{-7\lambda}\, d\lambda.$$

for $x \in \{0, 1, \ldots\}$.

©Nic Freeman, University of Sheffield, 2024.

Chapter 3

3.1 See `2_dist_sketching_solution.ipynb` or `2_dist_sketching_solution.Rmd`

3.2 (a) From Theorem 3.1.2 the posterior distribution has p.d.f.

$$\begin{aligned} f_{\Theta|_{\{X=2\}}} &= \frac{1}{Z} f_{\text{Exp}(\theta)}(2) f_{\text{Gamma}(2,3)}(\theta) \\ &= \frac{1}{Z} \frac{3^2}{\Gamma(2)} \theta e^{-2\theta} \theta e^{-3\theta} \\ &= \frac{1}{Z'} \theta^2 e^{-5\theta} \end{aligned}$$

for $\theta > 0$ and zero otherwise, where $\frac{1}{Z'} = \frac{1}{Z} \frac{3^2}{\Gamma(2)}$. We recognize $\Theta|_{\{X=2\}} \sim \text{Gamma}(3, 5)$.

(b) From (3.2) the sampling distribution has p.d.f.

$$\begin{aligned} f_X(x) &= \int_0^\infty f_{\text{Exp}(\theta)}(x) f_{\text{Gamma}(2,3)}(\theta)\, d\theta \\ &= \frac{3^2}{\Gamma(2)} \int_0^\infty \theta e^{-\theta x} \theta e^{-3\theta}\, d\theta \\ &= 9 \int_0^\infty \theta^2 e^{-\theta(x+3)}\, d\theta \\ &= 9 \frac{2}{(x+3)^3} \int_0^\infty \frac{(x+3)^3}{2} \theta^2 e^{-\theta(x+3)}\, d\theta \\ &= 9 \frac{2}{(x+3)^3} \int_0^\infty f_{\text{Gamma}(3,x+3)}(\theta)\, d\theta \\ &= \frac{18}{(x+3)^3} \end{aligned}$$

for $x > 0$ and zero otherwise.

From (3.5) and part (a), the corresponding predictive distribution has p.d.f.

$$\begin{aligned} f_{X'}(x) &= \int_0^\infty f_{\text{Exp}(\theta)}(x) f_{\text{Gamma}(3,5)}(\theta)\, d\theta \\ &= \frac{5^3}{\Gamma(3)} \int_0^\infty \theta e^{-\theta x} \theta^2 e^{-5\theta}\, d\theta \\ &= \frac{5^3}{2} \int_0^\infty \theta^3 e^{-\theta(x+5)}\, d\theta \\ &= \frac{5^3}{2} \frac{\Gamma(4)}{(x+5)^4} \int_0^\infty \frac{(x+5)^4}{\Gamma(4)} \theta^3 e^{-\theta(x+5)}\, d\theta \\ &= \frac{5^3}{2} \frac{6}{(x+5)^4} \int_0^\infty f_{\text{Gamma}(4,x+5)}(\theta)\, d\theta \\ &= \frac{375}{(x+5)^4} \end{aligned}$$

for $x > 0$ and zero otherwise.

(c) Using independence, the p.d.f. of the model family now becomes

$$f_{M_\theta}(x) = \prod_{i=1}^n \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_1^n x_i}.$$

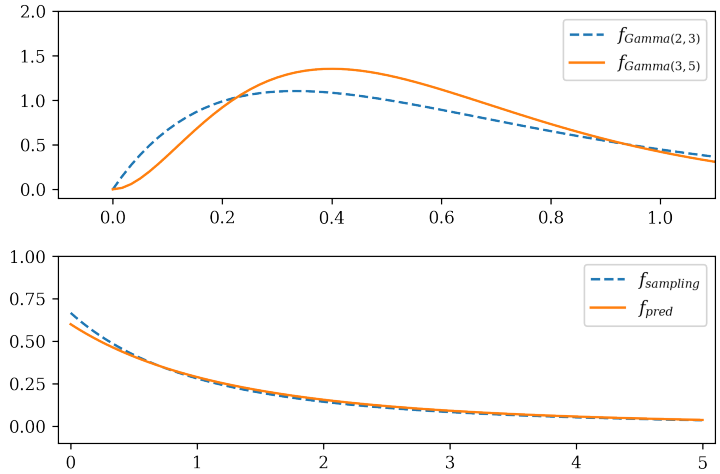
Let us write $z = \sum_1^n x_i$. From Theorem 3.1.2 we have

$$\begin{aligned} f_{\Theta|_{\{X=x\}}}(\theta) &= \frac{1}{Z} f_{M_\theta}(x) f_{\Gamma(2,3)}(\theta) \\ &= \frac{1}{Z} \frac{3^2}{\Gamma(2)} \theta^n e^{-\theta z} \theta e^{-3\theta} \\ &= \frac{1}{Z'} \theta^{n+1} e^{-\theta(3+z)} \end{aligned}$$

for $\theta > 0$ and zero otherwise. We recognize the Gamma$(n + 2, 3 + z)$ distribution.

©Nic Freeman, University of Sheffield, 2024.

3.3 We obtain:



3.4 (a) In order: Section 1.1, Section 1.2, Lemma 1.4.1, Lemma 1.5.1, equation (1.9), Lemma 1.6.1.

(b) The first part combines Definition 2.2.1 (discrete case) and Definition 3.1.1 (continuous case). The second part combines equations (2.4) and (3.2). The third part combines Theorems 2.4.1 and 3.1.2, as discussed at the end of Section 3.1. The fourth part combines equations (2.8) and (3.5).

3.5 *In this solution we will keep track of the normalizing constants. If you prefer to write them as $\frac{1}{Z}$ in the style of e.g.* (3.7) *and use Lemma 1.2.5 to recognize the distributions, or to use $\propto$ as introduced in Chapter 4, that is fine – but it is difficult to make that approach work for the predictive distribution in part (b)!*

In this question we need to be very careful with the limits of integrals. The Uniform$([0,\theta])$ p.d.f. is zero outside of $[0,\theta]$ and the Pareto(a,b) p.d.f. is zero outside of (b,∞). This matters particularly for the integrals in part (b).

(a) From Theorem 3.1.2 the posterior distribution has p.d.f.

$$\begin{aligned}
f_{\Theta|_{\{X=5\}}}(\theta) &= \frac{1}{\int_{\mathbb{R}} f_{\text{Uniform}([0,t])}(5) f_{\text{Pareto}(1,3)}(t)\,dt} f_{\text{Uniform}([0,\theta])}(\tfrac{1}{2}) f_{\text{Pareto}(1,3)}(\theta) \\
&= \frac{1}{\int_1^\infty \frac{1}{t} 3t^{-4}\,dt} \frac{1}{\theta} 3\theta^{-4} \\
&= \frac{1}{\int_1^\infty 3t^{-5}\,dt} 3\theta^{-5} \\
&= \frac{1}{3/4}\theta^{-5} \\
&= 4\theta^{-5}
\end{aligned}$$

for $\theta > 1$ and zero otherwise. We recognize the p.d.f. of the Pareto$(4,1)$ distribution. Note that to deduce the second line we used $f_{\text{Uniform}([0,\theta])}(\frac{1}{2}) = \frac{1}{\theta}$, which was true because $\frac{1}{2} < \theta$.

(b) From Theorem 3.1.2 the posterior distribution has p.d.f.

$$\begin{aligned}
f_{\Theta|_{\{X=5\}}}(\theta) &= \frac{1}{\int_{\mathbb{R}} f_{\text{Uniform}([0,t])}(5) f_{\text{Pareto}(1,3)}(t)\,dt} f_{\text{Uniform}([0,\theta])}(5) f_{\text{Pareto}(3,1)}(\theta) \\
&= \frac{1}{\int_1^\infty \mathbb{1}_{\{5\le t\}} \frac{1}{t} 3t^{-4}\,dt} \mathbb{1}_{\{5\le\theta\}} \frac{1}{\theta} 3\theta^{-4} \\
&= \frac{1}{\int_5^\infty 3t^{-5}\,dt} \mathbb{1}_{\{5\le\theta\}} 3\theta^{-5} \\
&= \frac{1}{3\frac{5^{-4}}{4}} \mathbb{1}_{\{5\le\theta\}}\theta^{-5} \\
&= \mathbb{1}_{\{5\le\theta\}} 5^4 4\theta^{-5}
\end{aligned}$$

We recognize the p.d.f. of the Pareto$(4,5)$ distribution.

©Nic Freeman, University of Sheffield, 2024.

From (3.5) the predictive distribution has p.d.f. given by

$$\begin{aligned}
f_{X'}(x) &= \int_5^\infty f_{\text{Uniform}([0,\theta])}(x) f_{\text{Pareto}(4,5)}(\theta)\, d\theta \\
&= \begin{cases} \int_x^\infty \frac{1}{\theta} 5^4 4\theta^{-5}\, d\theta & \text{for } x > 5 \\ \int_5^\infty \frac{1}{\theta} 5^4 4\theta^{-5}\, d\theta & \text{for } x \in [0,5] \end{cases} \\
&= \begin{cases} \int_x^\infty \mathbb{1}_{x\leq\theta} 5^4 4\theta^{-6}\, d\theta & \text{for } x > 5 \\ \int_5^\infty 5^4 4\theta^{-6}\, d\theta & \text{for } x \in [0,5] \end{cases} \\
&= \begin{cases} 5^4 4[\frac{\theta^{-5}}{-5}]_x^\infty & \text{for } x > 5 \\ 5^4 4[\frac{\theta^{-5}}{-5}]_5^\infty & \text{for } x \in [0,5] \end{cases} \\
&= \begin{cases} 5^4 4\frac{1}{5} x^{-5} & \text{for } x > 5 \\ 5^4 4\frac{1}{5} 5^{-5} & \text{for } x \in [0,5] \end{cases} \\
&= \begin{cases} 5^3 4x^{-5} & \text{for } x > 5 \\ \frac{4}{5^2} & \text{for } x \in [0,5] \end{cases}
\end{aligned}$$

for $x \geq 0$ and zero otherwise.

3.6 Noting that $\mathbb{P}[\Theta \in A] > 0]$, we will use Lemma 1.5.1. For $B \subseteq R_X$ we have

$$\begin{aligned}
\mathbb{P}[X|_{\{\Theta\in A\}} \in B] &= \frac{\mathbb{P}[X \in B, \Theta \in A]}{\mathbb{P}[\Theta \in A]} = \frac{1}{\mathbb{P}[\Theta \in A]} \int_B \int_A f_{M_\theta}(x) f_\Theta(\theta)\, d\theta\, dx \\
&= \int_B \int_A f_{M_\theta}(x) f_{\Theta|_{\{\Theta\in A\}}}(\theta)\, d\theta\, dx.
\end{aligned}$$

It follows from Definition 1.1.1 that $X|_{\{\Theta\in A\}}$ is a continuous random variable with p.d.f.

$$f_{X|_{\{\Theta\in A\}}}(x) = \int_A f_{M_\theta}(x) f_{\Theta|_{\{\Theta\in A\}}}(\theta)\, d\theta.$$

3.7 (a) We have

$$\int_{\mathbb{R}} f_{M'_\theta}(x)\, dx = \int_{\mathbb{R}} \int_{\mathbb{R}} f_{M_\theta}(x-y)\kappa(y)\, dy\, dx = \int_{\mathbb{R}} \kappa(y) \left(\int_{\mathbb{R}} f_{M_\theta}(x-y)\, dx \right) dy = \int_{\mathbb{R}} \kappa(y)\, dy = 1.$$

Here we used that f_{M_θ} is a p.d.f. which integrates to 1, and also our assumption that κ integrates to 1.

(b) By Theorem 3.1.2 we have $f_{\Theta|_{\{X=x\}}}(\theta) = \frac{1}{Z} f_{M_\theta}(x) f_\Theta(\theta)$ and also that

$$f_{\Theta|_{\{X'=x\}}}(\theta) = \frac{1}{Z'} f_{M'_\theta}(x) f_\Theta(\theta) = \frac{1}{Z'} \int_{\mathbb{R}} f_{M_\theta}(x-y)\kappa(y) f_\Theta(\theta)\, dy = \frac{Z}{Z'} \int_{\mathbb{R}} f_{\Theta|_{\{X=x-y\}}}(\theta)\kappa(y)\, dy$$

as required.

(c) In the case where $\kappa(x)$ is the p.d.f. of $\text{N}(0,1)$, the convolution applied to Model 1 is equivalent to adding a $\text{N}(0,1)$ random variable to the data, which gives Model 2. That is, $X' \overset{\text{d}}{=} X + \text{N}(0,1)$. Model 2 is therefore a version of Model 1 that is designed handle (additional) noise.

It helps to visualize things, which is left for you here: the effect of convolution on the probability density functions is to smooth them i.e. to spread out high peaks into lower and wider regions. Equation (3.10) says that the posterior density of (X', Θ) can be obtained by taking the posterior density of (X, Θ) and smoothing it in this way (with respect to the x coordinate).

Chapter 4

4.1 (a) (i) The posterior is

$$\text{N}\left(\frac{\frac{1}{4}(14.08) + \frac{0}{1}}{\frac{3}{4} + \frac{1}{1}}, \frac{1}{\frac{3}{4} + \frac{1}{1}} \right) \overset{\text{d}}{=} \text{N}(2.01, 0.76^2)$$

where we have rounded the parameters to two decimal places.

©Nic Freeman, University of Sheffield, 2024.

(ii) The p.d.f. of the sampling distribution is

$$f_X(x) = \int_{\mathbb{R}} f_{\mathrm{N}(\theta,2)}(x) f_{\mathrm{N}(0,1)}(\theta)\, d\theta.$$

The p.d.f. of the posterior distribution is

$$f_{X'}(x) = \int_{\mathbb{R}} f_{\mathrm{N}(\theta,2)}(x) f_{\mathrm{N}(2.01,0.76^2)}(\theta)\, d\theta.$$

(b) See `2_dist_sketching_solution.ipynb` and `2_dist_sketching_solution.Rmd`.

4.2 From Theorem 2.4.1 we have

$$\begin{aligned} f_{\Theta|_{\{X=x\}}}(\theta) &\propto \mathbb{P}[\mathrm{Geometric}(\theta)^{\otimes n} = x] f_{\mathrm{Beta}(\alpha,\beta)}(\theta) \\ &\propto \left(\prod_{i=1}^n \theta(1-\theta)^{x_i}\right) \frac{1}{\mathcal{B}(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^n (1-\theta)^{\sum_1^n x_i} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{\alpha+n-1}(1-\theta)^{\beta+\sum_1^n x_i - 1} \end{aligned}$$

for $\theta \in [0,1]$ and zero otherwise. Using Lemma 1.2.5, we recognize the $\mathrm{Beta}(\alpha+n, \beta+\sum_1^n x_i)$ distribution, as required.

4.3 From Theorem 2.4.1 we have

$$\begin{aligned} f_{\Theta|_{\{X=x\}}}(\theta) &\propto \mathbb{P}[\mathrm{Poisson}(\theta)^{\otimes n} = x] f_{\mathrm{Gamma}(\alpha,\beta)}(\theta) \\ &\propto \left(\prod_{i=1}^n \frac{\theta^{x_i} e^{-\theta}}{x_i!}\right) \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \\ &\propto \theta^{\sum_1^n x_i} e^{-n\theta} \theta^{\alpha-1} e^{-\beta\theta} \\ &\propto \theta^{\alpha+\sum_1^n x_i - 1} e^{-\theta(\beta+n)} \end{aligned}$$

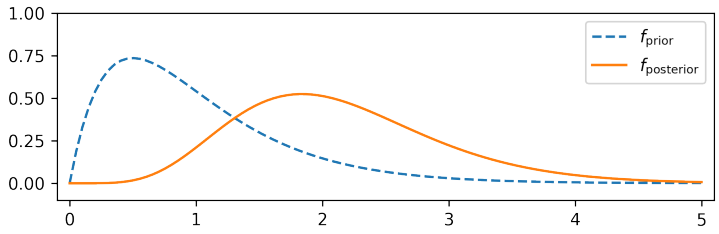
for $\theta > 0$ and zero otherwise. Using Lemma 1.2.5, we recognize the $\mathrm{Gamma}(\alpha+\sum_1^n x_i, \beta+n)$ distribution, as required.

4.4 From Theorem 3.1.2 we have

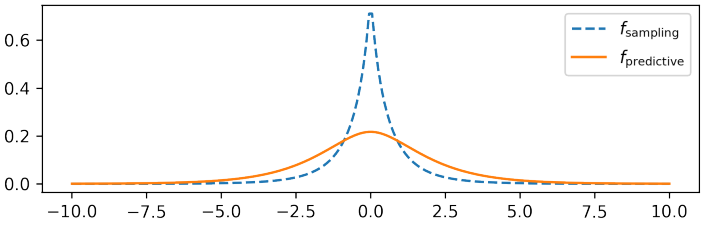
$$\begin{aligned} f_{\Theta|_{\{X=x\}}}(\tau) &\propto f_{\mathrm{N}(\mu,\frac{1}{\tau})^{\otimes n}}(x) f_{\mathrm{Gamma}(\alpha,\beta)}(\tau) \\ &\propto \left(\prod_{i=1}^n \frac{\sqrt{\tau}}{\sqrt{2\pi}} e^{-\frac{\tau(x_i-\mu)^2}{2}}\right) \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau} \\ &\propto \tau^{\frac{n}{2}} \exp\left(-\frac{\tau}{2}\sum_1^n (x_i-\mu)^2\right) \tau^{\alpha-1} e^{-\beta\tau} \\ &\propto \tau^{\alpha+\frac{n}{2}-1} \exp\left[-\tau\left(\beta + \frac{1}{2}\sum_1^n (x_i-\mu)^2\right)\right]. \end{aligned}$$

Using Lemma 1.2.5, we recognize the $\mathrm{Gamma}(\alpha+\frac{n}{2}, \beta+\frac{1}{2}\sum_1^n (x_i-\mu)^2)$ distribution, as required.

4.5 We obtain



©Nic Freeman, University of Sheffield, 2024.



4.6 (a) The posterior is $N(2.013, 0.36)$.

(b) Writing the code is left for you. The result will be the same as in part (a).

(c) We have seen that the Bayesian updates here can be done all at once, or piece by piece, and will give the same results. Checking the formulae for the conjugate priors, in every other case covered in this chapter it is obvious that this will be the case – only in Lemma 4.2.2 are the update formulae complicated enough that it is not obvious from the formulae.

*In fact, this principle holds with or without conjugate pairs, as we will see in Exercise **6.7**.*

4.7 From Theorem 3.1.2 we have

$$\begin{aligned} f_{\Theta|_{\{X=x\}}}(\theta) &\propto f_{\text{Weibull}(\theta,\beta)^{\otimes n}}(x) f_{\text{IGamma}(a,b)}(\theta) \\ &\propto \left(\prod_{i=1}^{n} \theta\beta(x_i)^{\beta-1} e^{-\theta x_i^{\beta}} \right) \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \\ &\propto \theta^n e^{-\theta \sum_1^n x_i^{\beta}} \theta^{a-1} e^{-b\theta} \\ &\propto \theta^{a+n-1} e^{-\theta(b+\sum_1^n x_i^{\beta})}. \end{aligned}$$

Using Lemma 1.2.5, we recognize the $\Gamma(\alpha+n, \beta+\sum_1^n x_i^k)$ distribution, as required.

4.8 Omitted.

4.9 (a) Taking $C = 1$ in Definition 4.1.1 gives $f \propto f$.

(b) If $f(x) = Cg(x)$ then $g(x) = \frac{1}{C} f(x)$. Note that Definition 4.1.1 gives $C > 0$, so $\frac{1}{C} > 0$.

(c) If $f(x) = Cg(x)$ and $g(x) = C'h(x)$ then $f(x) = CC'h(x)$. Note that Definition 4.1.1 gives $C, C' > 0$, so $CC' > 0$.

Chapter 5

5.1 The data from Census 2021 is as follows.

Age band	Population	Proportion (2dp)
10+	60096227	0.89
20+	52089688	0.77
30+	43661735	0.65
40+	34458842	0.51
50+	26028478	0.39
60+	16814195	0.25
70+	9316631	0.14
80+	3399106	0.05
90+	609904	0.01

The total population was 67596281.

The point of this question that you will (probably) find it more difficult to give accurate estimates for events that have smaller probabilities.

5.2 This is up to you!

5.3 For the model family $(M_\lambda)_{\lambda\in(0,\infty)}$ in which $M_\lambda \sim \text{Poisson}(\lambda)$ we have

$$L_{M_\lambda}(x) = \begin{cases} \dfrac{\lambda^x e^{-\lambda}}{x!} & \text{for } \lambda > 0 \\ 0 & \text{otherwise} \end{cases}$$

©Nic Freeman, University of Sheffield, 2024.

where $x \in \{0, 1, \ldots\}$. Hence $\frac{d}{d\lambda} \log(L_{M_\lambda}(x)) = \frac{d}{d\lambda}(x \log(\lambda) - \lambda - x!) = \frac{x}{\lambda} - 1 = \frac{x-\lambda}{\lambda}$. For $\lambda > 0$, the density function of the Jeffrey's prior is given by

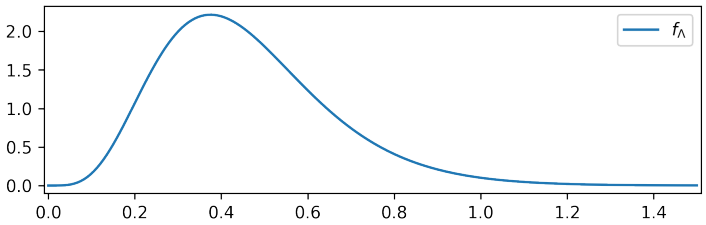
$$\begin{aligned} f_\Lambda(\lambda) &\propto \mathbb{E}\left[\left(\frac{d}{d\theta} \log(L_{M_\theta}(X))\right)^2\right]^{1/2} \\ &\propto \mathbb{E}\left[\left(\frac{X-\lambda}{\lambda}\right)^2\right]^{1/2} \\ &\propto \frac{1}{\lambda}\mathbb{E}\left[(X-\lambda)^2\right]^{1/2} \\ &\propto \frac{1}{\lambda}\operatorname{var}(X)^{1/2} \\ &\propto \frac{1}{\lambda}\lambda^{1/2} \\ &\propto \frac{1}{\lambda^{1/2}}. \end{aligned}$$

Noting that $\int_0^\infty \frac{1}{\lambda^{1/2}}\, d\lambda = \infty$, this is an improper prior.

5.4 Our model here is $M_\lambda = \text{Poisson}(\lambda)^{\otimes 12}$. From Theorem 2.4.1 we have

$$\begin{aligned} f_{\Lambda|_{\{X=x\}}}(x) &\propto \left(\prod_{i=1}^{1} 2\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right)\frac{1}{\sqrt{\lambda}} \\ &\propto \lambda^{\sum_1^{12} x_i - \frac{1}{2}} e^{-12\lambda}. \end{aligned}$$

Using Lemma 1.2.5 we recognize the Gamma$(\frac{1}{2} + \sum_1^n x_i, n)$ distribution with $n = 12$. This is a proper distribution for all values of x, which answers part (b). For part (a) we have $\sum_1^{12} x_i = 5$, so we obtain $\Lambda|_{\{X=x\}} \sim \text{Gamma}(\frac{11}{2}, 12)$. A sketch looks like



5.5 (a) From Theorem 3.1.2 we have

$$f_{\Theta|_{\{X=x\}}}(\theta) \propto f_{\text{Uniform}(0,\theta)}(x) f_{\text{Exp}(1)}(\theta) \propto \begin{cases} \frac{1}{\theta}\theta e^{-\theta x} & \text{for } \theta > x > 0, \\ 0 & \text{otherwise,} \end{cases} \propto \begin{cases} e^{-\theta x} & \text{for } \theta > x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The distribution Uniform$(0, \theta)$ only generates values in $(0, \theta)$, so in order to generate the data x it must have $x \in (0, \theta)$. For this reason our posterior places zero weight on $\theta < x$. (The boundary $x = \theta$ has probability zero, so it does not matter what happens there.)

(b) The posterior is not well defined in this case. Formally the condition $x \in R$ of Theorem 3.1.2 is not satisfied. From a more practical point of view, we have failed to account for Cromwell's rule. Our prior density is zero outside of $\theta \in (1, 2)$ but our model makes sense for all $\theta > 0$, and to generate the data $x = 3$ we would need to have $\theta > 3$.

5.6 We argue by contradiction: suppose such a U does exist. Take $c = 1$ and $[a, b] = [n, n + 1]$ and we obtain that $\mathbb{P}[U \in [n, n + 1]] = \mathbb{P}[U \in [n + 1, n + 2]]$. By a trivial induction we have $\mathbb{P}[U \in [0, 1]] = \mathbb{P}[U \in [1, 2]] = \mathbb{P}[U \in [2, 3]] = \ldots$, but then

$$1 = \mathbb{P}[U \in [0, \infty)] = \sum_{n=0}^{\infty} \mathbb{P}[U \in [n, n + 1]] = \sum_{n=0}^{\infty} \mathbb{P}[U \in [0, 1]].$$

This a contradiction: the right hand side is either 0 (if $\mathbb{P}[U \in [0, 1]] = 0$) or equal to ∞ (if $\mathbb{P}[U \in [0, 1]] > 0$).

©Nic Freeman, University of Sheffield, 2024.

5.7 Bob chooses his prior to be $h(\Theta)$, where Θ is Alice's prior with p.d.f. f_1. As h is strictly monotone increasing and differentiable, this means that Bob's prior has p.d.f.

$$f_2(\theta) = \frac{dh^{-1}}{d\theta} f_1\left(h^{-1}(\theta)\right).$$

Take Alice's sampling distribution and substitute $\theta = h(\lambda)$ to obtain

$$\begin{aligned} f_{X_1}(x) &= \int_\Pi f_{M_\theta}(x) f_1(\theta)\, d\theta \\ &= \int_\Pi f_{M_{h(\lambda)}}(x) f_1\left(h^{-1}(\lambda)\right) \frac{dh^{-1}}{d\lambda}\, d\lambda \\ &= \int_\Pi f_{M_{h(\lambda)}}(x) f_2(\lambda)\, d\lambda \\ &= f_{X_2}(x) \end{aligned}$$

as required.

5.8 By independence we have $L_{M_\theta^{\otimes n}}(x) = \prod_{i=1}^n L_{M_\theta}(x_i)$, hence $\log L_{M_\theta^{\otimes n}}(x) = \sum_1^n \log L_{M_\theta}(x_i)$. Taking $X = (X_i) \sim M_\theta^{\otimes n}$ so that $X_i \sim M_\theta$ for all i, we have

$$\begin{aligned} \mathbb{E}\left[-\frac{d}{d\theta^2} \log L_{M_\theta^{\otimes n}}(X)\right] &= \mathbb{E}\left[-\frac{d}{d\theta^2} \sum_1^n \log L_{M_\theta}(X_i)\right] \\ &= \sum_{i=1}^n \mathbb{E}\left[-\frac{d}{d\theta^2} \log L_{M_\theta}(X_i)\right]. \end{aligned}$$

Hence by (5.3)

$$f_{M_\theta^{\otimes n}}(\theta) \propto \sum_{i=1}^n f_{M_\theta}(\theta) \propto n f_{M_\theta}(\theta) \propto f_{M_\theta}(\theta)$$

as required.

Chapter 6

6.1 We have $f_{\text{Gamma}(\alpha,\beta)}(x) \propto x^{\alpha-1}e^{-\beta x}$ when $x > 0$ and zero otherwise. Differentiating, we have

$$(\alpha-1)x^{\alpha-2}e^{-\beta x} + x^{\alpha-1}(-\beta)e^{-\beta x} = x^{\alpha-2}e^{-\beta x}(\alpha - 1 - \beta x).$$

This takes the value zero when, and only when, $\alpha - 1 - \beta x = 0$, which gives $x = \frac{\alpha-1}{\beta}$. If $\alpha \geq 1$ then this value is within the range of Gamma(α, β). Since $f_{\text{Gamma}(\alpha,\beta)}(0) = 0$ and $\lim_{x\to\infty} f_{\text{Gamma}(\alpha,\beta)}(x) = 0$, and there is only one turning point, that turning point must be a global maximum. Hence it is also the mode, given by $\frac{\alpha-1}{\beta}$.

If $\alpha \in (0, 1)$ then from (C) we have that $f_{\text{Gamma}(\alpha,\beta)}(x)$ has negative derivative for all $x > 0$. Hence it is a decreasing function, and the maximum will occur at $x = 0$. So the mode is zero, when $\alpha \in (0, 1)$.

6.2 (a) Lemma 1.6.1.

(b) Lemma 4.1.5.

(c) If $\lambda \sim$ Gamma(α, β) and $x|\lambda \sim \text{Exp}(\lambda)^{\otimes n}$ then $\lambda|x \sim$ Beta$(\alpha + n, \beta + \sum_1^n x_i)$.

(d) If $(\mu, \tau) \sim$ NGamma(m, p, a, b) then $\tau \sim$ Gamma(a, b) and $\mu|\tau \sim \text{N}(m, \frac{1}{p\tau})$.

6.3 (a) If $X \sim N(0, 1)$ then $X|_{\{X>0\}} \overset{\text{d}}{=} |X|$. This is the result of Exercise **1.7**, which is closely related to Example 1.4.3.

(b) If X and Y are independent random variables then $X|_{\{Y=y\}} \overset{\text{d}}{=} X$. This is true if the conditioning is well defined, as discussed (in more general terms) at the start of Section 1.5.

6.4 (a) We have $f_{\text{NegBin}(m,\theta)}(x_i) \propto \theta^m(1-\theta)^{x_i}$, for $x_i \in \{0, 1, \dots, \}$. Hence

$$f(x|\theta) = f_{\text{NegBin}(m,\theta)^{\otimes n}}(x) \propto \prod_{i=1}^n \theta^m(1-\theta)^{x_i} \propto \theta^{mn}(1-\theta)^{\sum_1^n x_i}.$$

©Nic Freeman, University of Sheffield, 2024.

(b) From Theorem 3.1.2 we have

$$\begin{aligned} f(\theta|x) &\propto f_{\mathrm{NegBin}(m,\theta)^{\otimes n}}(x) f_{\mathrm{Beta}(\alpha,\beta)}(\theta) \\ &\propto \theta^{mn}(1-\theta)^{\sum_1^n x_i}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{\alpha+mn-1}(1-\theta)^{\beta+\sum_1^n x_i-1}. \end{aligned}$$

By Lemma 1.2.5 we recognize $\theta|x \sim \mathrm{Beta}(\alpha^*, \beta^*)$ with $\alpha^* = \alpha + mn$ and $\beta^* = \beta + \sum_1^n x_i$.

(c) (i) The reference prior is given by

$$\begin{aligned} f(\theta) &\propto \mathbb{E}\left[-\frac{d^2}{d\theta^2}\log L_{\mathrm{NegBin}(m,\theta)}(X)\right]^{1/2} \\ &\propto \mathbb{E}\left[-\frac{d^2}{d\theta^2}\log\left(\theta^{mn}(1-\theta)^{\sum_1^n X_i}\right)\right]^{1/2} \\ &\propto \mathbb{E}\left[-\frac{d^2}{d\theta^2}\left(mn\log\theta + \sum_1^n X_i\log(1-\theta)\right)\right]^{1/2} \\ &\propto \mathbb{E}\left[\frac{mn}{\theta^2} + \frac{\sum_1^n X_i}{(1-\theta)^2}\right]^{1/2} \\ &\propto \left(\frac{mn}{\theta^2} + \frac{mn(1-\theta)}{\theta}\frac{1}{(1-\theta)^2}\right)^{1/2} \end{aligned}$$

where we use that $X_i \sim \mathrm{NegBin}(m,\theta)$ has mean $\frac{m(1-\theta)}{\theta}$. Hence

$$f(\theta) \propto \left(\frac{mn(1-\theta)+mn\theta}{\theta^2(1-\theta)}\right)^{1/2} \propto \theta^{-1}(1-\theta)^{-1/2}.$$

(ii) $f(\theta) \propto \theta^{-1}(1-\theta)^{-1/2}$ not define a proper distribution. To see this,

$$\int_0^{\frac{1}{2}} \theta^{-1}(1-\theta)^{-1/2}\,d\theta \geq \int_0^{\frac{1}{2}} \theta^{-1}(1/2)^{-1/2}\,d\theta = \infty.$$

(iii) From Theorem 3.1.2 the prior is given by

$$\begin{aligned} f(\theta|x) &\propto \theta^{mn}(1-\theta)^{\sum_1^n x_i}\theta^{-1}(1-\theta)^{-1/2} \\ &\propto \theta^{mn-1}(1-\theta)^{\sum_1^n x_i-\frac{1}{2}}. \end{aligned}$$

Using Lemma 1.2.5 we recognize $\theta|x \sim \mathrm{Beta}(mn, \sum_1^n x_i + \frac{1}{2})$.

6.5 (a) By Theorem 3.1.2 the posterior distribution has p.d.f.

$$\begin{aligned} f(\mu,\tau|x) &\propto \left(\prod_{i=1}^n f_{\mathbb{N}(\mu,\frac{1}{\tau})}(x_i)\right)\frac{1}{\tau} \\ &\propto \left(\prod_{i=1}^n \frac{1}{\sqrt{\tau}}e^{-\frac{1}{2}\tau(x_i-\mu)^2}\right)\frac{1}{\tau} \\ &\propto \tau^{\frac{n}{2}-1}\exp\left(-\frac{\tau}{2}\sum_1^n(x_i-\mu)^2\right) \end{aligned}$$

where $s^2 = \frac{1}{n}\sum_1^n(x_i-\mu)^2$.

(b) To find the marginal distribution of τ we must integrate over μ, giving

$$\begin{aligned} f(\tau|x) &\propto \int_{\mathbb{R}} \tau^{\frac{n}{2}-1}\exp\left(-\frac{\tau}{2}\sum_1^n(x_i-\mu)^2\right)d\mu \\ &\propto \int_{\mathbb{R}} \tau^{\frac{n}{2}-1}\exp\left(-\frac{\tau}{2}\left(ns^2+(\bar{x}-\mu)^2\right)\right)d\mu \\ &\propto \tau^{\frac{n}{2}-1-\frac{1}{2}}e^{-\frac{1}{2}\tau ns^2}\int_{\mathbb{R}} \tau^{1/2}\exp\left(-\frac{\tau}{2}(\bar{x}-\mu)^2\right)d\mu \end{aligned}$$

©Nic Freeman, University of Sheffield, 2024.

$$\propto \tau^{\frac{n-1}{2}-1} e^{-\frac{1}{2}\tau n s^2} \int_{\mathbb{R}} f_{\mathrm{N}(\bar{x},\frac{1}{\tau})(\mu)}\, d\mu$$
$$\propto \tau^{\frac{n-1}{2}-1} e^{-\frac{1}{2}\tau n s^2}.$$

Using Lemma 1.2.5 for $n \geq 2$ we recognize $\tau|x \sim \mathrm{Gamma}(a^*, b^*)$ where $a^* = \frac{n-1}{2}$ and $b^* = \frac{1}{2}ns^2$. If $n = 1$ then this does not correspond to a Gamma distribution, because in this case $a^* = 0$ and the Gamma distribution requires parameters in $(0, \infty)$.

We have

$$\int_0^\infty \int_{\mathbb{R}} f(\mu, \tau|x)\, d\mu\, d\tau = \int_0^\infty f(\tau|x)\, d\tau$$
$$\propto \int_0^\infty f(\tau|x)\, d\tau,$$

which is finite if $n \geq 2$ because the Gamma distribution is proper. If $n = 1$ then we have

$$\int_0^\infty \int_{\mathbb{R}} f(\mu, \tau|x)\, d\mu\, d\tau = \int_0^\infty \tau^{-1} e^{-\frac{1}{2}\tau s^2}\, d\tau \geq e^{-\frac{1}{2}s^n} \int_0^1 \frac{1}{\tau}\, d\tau = \infty. \tag{C.3}$$

Here we use that $e^{-\frac{1}{2}\tau s^2} \geq e^{-\frac{1}{2}s^n}$ for $\tau \in [0, 1]$. Hence $f(\mu, \tau|x)$ defines an improper distribution when $n = 1$.

6.6 (a) We have $f_{\Theta_i}(\theta) \geq 0$ and $\alpha, \beta \geq 0$ so also $f_\Theta(\theta) \geq 0$. Also,

$$\int_{\mathbb{R}} f_\Theta(\theta)\, d\theta = \alpha \int_{\mathbb{R}} f_{\Theta_1}(\theta)\, d\theta + \beta \int_{\mathbb{R}} f_{\Theta_2}(\theta)\, d\theta = \alpha(1) + \beta(1) = 1.$$

(b) By Theorem 3.1.2 we have

$$f_{\Theta|_{\{X=x\}}}(\theta) = \frac{f_{M_\theta}(x) f_\Theta(\theta)}{\int_{\mathbb{R}^n} f_{M_\theta}(x) f_\Theta(\theta)\, dx}$$
$$= \frac{\alpha f_{M_\theta}(x) f_{\Theta_1}(\theta) + \beta f_{M_\theta}(x) f_{\Theta_2}(\theta)}{\int_{\mathbb{R}^n} f_{M_\theta}(x) f_{\Theta_1}(\theta)\, dx + \int_{\mathbb{R}^n} f_{M_\theta}(x) f_{\Theta_2}(\theta)\, dx}$$
$$= \alpha' f_{\Theta_1|_{\{X_1=x\}}}(\theta) + \beta' f_{\Theta_1|_{\{X_1=x\}}}(\theta)$$

where

$$\alpha' = \frac{\alpha \int_{\mathbb{R}^n} f_{M_\theta}(x) f_{\Theta_1}(\theta)\, d\theta}{\alpha \int_{\mathbb{R}^n} f_{M_\theta}(x) f_{\Theta_1}(\theta)\, dx + \beta \int_{\mathbb{R}^n} f_{M_\theta}(x) f_{\Theta_2}(\theta)\, dx} = \frac{\alpha Z_1}{\alpha Z_1 + \beta Z_2}$$
$$\beta = \frac{\beta \int_{\mathbb{R}^n} f_{M_\theta}(x) f_{\Theta_2}(\theta)\, d\theta}{\alpha \int_{\mathbb{R}^n} f_{M_\theta}(x) f_{\Theta_1}(\theta)\, dx + \beta \int_{\mathbb{R}^n} f_{M_\theta}(x) f_{\Theta_2}(\theta)\, dx} = \frac{\beta Z_2}{\alpha Z_1 + \beta Z_2}$$

where Z_1 and Z_2 are the normalizing constants (from Theorem 3.1.2) for $f_{\Theta_1|_{\{X_1=x\}}}$ and $f_{\Theta_2|_{\{X_1=x\}}}$ respectively, as required.

(c) To cover discrete Bayesian models, instead of probability density functions f_{M_θ} in (b) we can use probability mass functions p_{M_θ}. We then need Theorem 2.4.1 in place of Theorem 3.1.2, but the argument is otherwise the same.

We could also cover both cases at once by using likelihood functions and (6.2).

6.7 (a) (i) From the combined version of Bayes rule (6.2) we have

$$f_{\Theta|_{\{X=x\}}}(\theta) \propto L_{M_\theta^{\otimes n}}(x) f_\Theta(\theta).$$

Applying (6.2) twice, we obtain

$$f_{\Theta|_{\{X_1=x(1)\}}}(\theta) \propto L_{M_\theta^{\otimes n_1}}(x(1)) f_\Theta(\theta)$$
$$f_{(\Theta|_{\{X_1=x(1)\}})|_{\{X_2=x(2)\}}}(\theta) \propto L_{M_\theta^{\otimes n_2}}(x(2)) L_{M_\theta^{\otimes n_1}}(x(1)) f_\Theta(\theta)$$

We note that by independence

$$L_{M_\theta^{\otimes n_2}}(x(2)) L_{M_\theta^{\otimes n_1}}(x(1)) = \left(\prod_{i=1}^{n_1} L_{M_\theta}(x_i)\right)\left(\prod_{i=n_1+1}^{n_2} L_{M_\theta}(x_i)\right) = \left(\prod_{i=1}^{n_2} L_{M_\theta}(x_i)\right) = L_{M_\theta^{\otimes n}}(x).$$

Hence $f_{(\Theta|_{\{X_1=x(1)\}})|_{\{X_2=x(2)\}}}(\theta) \propto f_{\Theta|_{\{X=x\}}}(\theta)$, which by Lemma 1.2.5 implies that $(\Theta|_{\{X_1=x(1)\}})|_{\{X_2=x(2)\}} \overset{\mathrm{d}}{=} \Theta|_{\{X=x\}}$ as required.

©Nic Freeman, University of Sheffield, 2024.

(ii) Applying a trivial induction to the result in part (a) we have shown that, in general for independent data, performing Bayesian updates in individual steps for each data point (or combinations of datapoints) will give the same results as performing one Bayesian update with all our data at once. This implies the result of Exercise **4.6**.

(b) From Bayes rule we have

$$f(\theta|x) \propto f(x|\theta)f(\theta)$$
$$f(\theta|x(1)) \propto f(x(1)|\theta)f(\theta)$$
$$f(\theta|x(1),x(2)) \propto f(x(2)|\theta)f(x(1)|\theta)f(\theta).$$

By independence (or more strictly, by conditional independence of $x(1)$ and $x(2)$ given θ) we have

$$f(x|\theta) = f(x(1),x(2)|\theta) \propto f(x(1)|\theta)f(x(2)|\theta)$$

hence $f(\theta|x) \propto f(\theta|x(1),x(2))$. By Lemma 1.2.5 we have $\theta|x \overset{\mathrm{d}}{=} \theta|x(1),x(2)$.

Chapter 7

7.1 f_X can be reasonably approximated by its mode. The median and mean would not be particularly bad choices, but the mode is best because the long right-hand tail (that does not contain much mass) will pull the median and mean slightly rightwards, away from the from region of highest density.

f_Y is tricky, because now the right hand tail contains substantial mass. It would be better not to approximate it with a point estimate, but if we had to do so then the median or mean would be reasonable choices, depending on the context. Alongside our point estimate, we should try to make sure the right-hand skew of the distribution is communicated in some way. The mode will be far below most of the mass of the distribution, so is a bad choice here.

There is no reasonable way to approximate f_Z with a point estimate. We could not capture the key feature of the distribution: that it has two approximately evenly sized peaks in different regions.

7.2 (a) We find numerically that the prior and posterior odds ratios are

$$\frac{\mathbb{P}[\mathrm{Beta}(2,8) > 0.2]}{\mathbb{P}[\mathrm{Beta}(2,8) \leq 0.2]} = 0.77 \qquad \text{and} \qquad \frac{\mathbb{P}[\mathrm{Beta}(11,19) > 0.2]}{\mathbb{P}[\mathrm{Beta}(11,19) \leq 0.2]} = 49.75$$

to two decimal places. The Bayes factor is 64.30.

(b) Note that $L_{\mathrm{Bin}(m,p)^{\otimes n}}(x) \propto p^x(1-p)^{m-x}$ so $\log L_{\mathrm{Bin}(m,p)}(x) = x\log p + (m-x)\log(1-p) = x\log p + m\log(1-p) - x\log(1-p)$. Hence $\frac{d}{dp}\log L_{\mathrm{Bin}(m,p)}(x) = \frac{x}{p} + \frac{1-x}{1-p}$, which is equal to $\frac{d}{dp}\log L_{\mathrm{Bernoulli}(p)}(x)$ as obtained in Example 5.3.4. Hence the same calculation as in Example 5.3.4 obtains the same reference prior.

We find numerically that the prior and posterior odds ratios are

$$\frac{\mathbb{P}[\mathrm{Beta}(\frac{1}{2},\frac{1}{2}) > 0.2]}{\mathbb{P}[\mathrm{Beta}(\frac{1}{2},\frac{1}{2}) \leq 0.2]} = 2.39 \qquad \text{and} \qquad \frac{\mathbb{P}[\mathrm{Beta}(\frac{1}{2}+9,\frac{1}{2}+11) > 0.2]}{\mathbb{P}[\mathrm{Beta}(\frac{1}{2}+9,\frac{1}{2}+11) \leq 0.2]} = 191.15$$

to two decimal places. The Bayes factor is 80.05.

The Bayes factor has not changed much, even though the odds ratios are quite different. For both choices of prior it suggests strong evidence for H_0 over H_1.

(c) The regulator will be interested by both sets of analysis. In particular, by the fact that both Bayes factors (using the informative prior elicited from the scientist and the uninformative reference prior) point towards the hypothesis $\theta > 0.2$, suggests that the analysis is robust i.e. is not overly sensitive to the particular methodology used.

This is a highly stylized example. Medical trials tend to be complex experiments with multiple subgroups, typically with outcomes that are not easily reducible to success vs. failure. The process of deciding what statistics will be reported is often done in negotiation with the regulator, before the trial begins.

7.3 (a) The reference density $f(\lambda)$ is not proper, so we cannot calculate the probabilities that $\lambda \in [0,2)$ or $\lambda \in [2,\infty)$ with respect to this density. The prior odds are not well-defined in this situation.

(b) The data given has $n = 40$ and $\sum_1^n x_i = 109$. Hence the posterior is $\lambda|x \sim \mathrm{Gamma}(\frac{5}{4}+18, \frac{1}{5}+8) \overset{\mathrm{d}}{=} \mathrm{Gamma}(23.5, 8.2)$. We find numerically that the prior and posterior odd ratios are

$$\frac{\mathbb{P}[\mathrm{Gamma}(\frac{5}{4},\frac{1}{5}) \geq 2]}{\mathbb{P}[\mathrm{Gamma}(\frac{5}{4},\frac{1}{5}) < 2]} = 3.42 \qquad \text{and} \qquad \frac{\mathbb{P}[\mathrm{Gamma}(23.5,8.2) \geq 2]}{\mathbb{P}[\mathrm{Gamma}(23.5,8.2) < 2]} = 1093.84$$

©Nic Freeman, University of Sheffield, 2024.

to two decimal places. The Bayes factor is 319.75.

We have (very) strong evidence for H_0 over H_1. A Poisson model is known to be reasonable, at least over large enough intervals of time, for large earthquakes. Assuming that we believe this model, we have strong evidence that Japan will, on average, experience two or more earthquakes of magnitude above 7.5 every year.

In case this sounds like unreasonably many earthquakes: earthquakes can occur deep underground, as well as offshore, and in such a case you may not hear much about them.

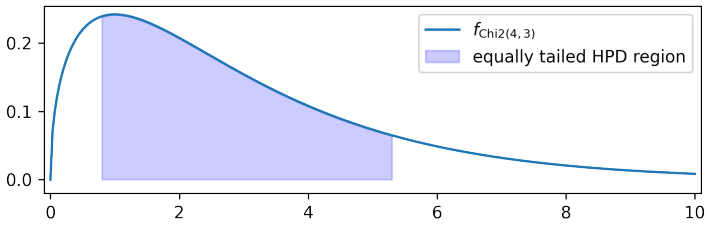
(c) For our new hypothesis $H_0 : \lambda \geq 3$, we find numerically that the prior and posterior odd ratios are

$$\frac{\mathbb{P}[\text{Gamma}(\frac{5}{4}, \frac{1}{5}) \geq 3]}{\mathbb{P}[\text{Gamma}(\frac{5}{4}, \frac{1}{5}) < 3]} = 1.95 \qquad \text{and} \qquad \frac{\mathbb{P}[\text{Gamma}(23.5, 8.2) \geq 3]}{\mathbb{P}[\text{Gamma}(23.5, 8.2) < 3]} = 0.19$$

to two decimal places. The Bayes factor is 0.09.

There is no evidence here to favour H_0 over the opposite hypothesis $H_1 : \lambda < 3$. In fact, swapping H_0 and H_1 will mean that the Bayes factor becomes $1/B$, which in this case is $1/0.09 = 10.14$, meaning that we have strong evidence to favour H_1 over H_0.

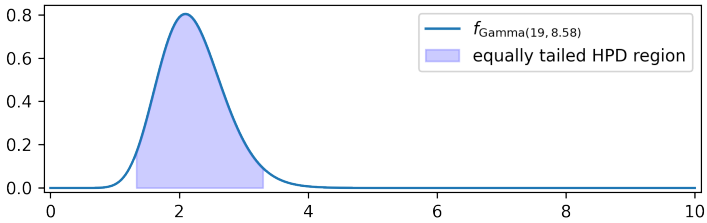
7.4 You should obtain $a = 0.80$ and $b = 5.32$ to two decimal places, and the following figure.



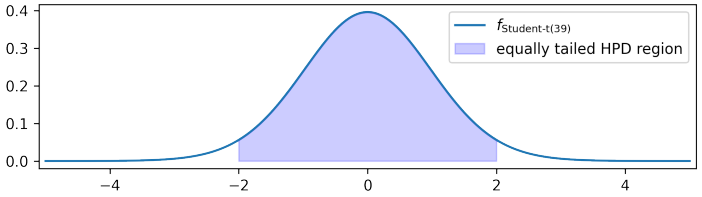
7.5 (a) I will spare you the details, but I have conducted an elicitation procedure on my wife who has (somewhat reluctantly) supplied us with the prior distribution $\text{Exp}(\frac{1}{1.7})$.

(b) The dataset has $n = 8$ and $\sum_1^n x_i = 18$. Hence the posterior distribution is $\text{Gamma}(1 + 18, \frac{1}{1.7} + 8) = \text{Gamma}(19, 8.59)$ with parameters to two decimal places.

(c) An equally tailed 95% HPD region is $[1.33, 3.31]$, to two decimal places, and looks like



7.6 (a) From the dataset we have $n = 40$, $\bar{x} = 7.55$ and $s^2 = 0.51$, to two decimal places. Hence $t|x \sim$ Student-t(39) with a 95% HPD region $[-2.02, 2.02]$,



The relationship between $\mu|x$ and $t|x$ is that $\mu|x = \bar{x} + (t|x)\frac{\sqrt{n}}{s}$. Note that the distribution of $t|x$ is symmetric about 0, hence the distribution of $\mu|x = \bar{x} + (t|x)\frac{\sqrt{n}}{s}$ will also be symmetric about its mean, and the mean of $\mu|x$ will be $\bar{x}$. An equally tailed 95% HPD region is given by $[\bar{x} - 2.02\frac{\sqrt{n}}{s}, \bar{x} + 2.02\frac{\sqrt{n}}{s}]$. Putting in $\bar{x}$, n and s, and this comes out as $[7.33, 7.78]$ to two decimal places.

©Nic Freeman, University of Sheffield, 2024.

(b) The posterior density function is given by

$$\begin{aligned} f(\mu,\phi) &= \left(\prod_{i=1}^{n} f_{\mathrm{N}(\mu,\phi)}(x_i)\right)\frac{1}{\phi} \\ &\propto \frac{1}{\phi^{n/2+1}}\exp\left(-\frac{1}{2\phi}\sum_{1}^{n}(x_i-\mu)^2\right) \\ &\propto \frac{1}{\phi^{n/2+1}}\exp\left(-\frac{1}{2\phi}\left(ns^2+n(\bar{x}-\mu)^2\right)\right) \end{aligned}$$

where we have used (4.10) and the notation of that identity, $\bar{x}=\sum_1^n x_i$ and $s^2=\frac{1}{n}\sum_1^n(x_i-\bar{x})^2$. Hence the marginal density $f(\mu|x)$ satisfies

$$f(\mu|x)\propto\int_0^\infty \frac{1}{\phi^{n/2+1}}\exp\left(-\frac{1}{2\phi}\left(ns^2+n(\bar{x}-\mu)^2\right)\right)d\phi.$$

To compute this integral we make the substitution $\psi=\frac{1}{2\phi}(ns^2+n(\bar{x}-\mu)^2)$. We have $\frac{d\psi}{d\phi}=\frac{-1}{2\phi^2}(ns^2+n(\bar{x}-\mu)^2)$, and hence

$$\begin{aligned} f(\mu|x) &\propto \int_0^\infty \frac{1}{\phi^{n/2+1}}e^{-\psi}\frac{2\phi^2}{ns^2+n(\bar{x}-\mu)^2}\,d\psi \\ &\propto \frac{1}{(ns^2+n(\bar{x}-\mu)^2)^{n/2}}\int_0^\infty \psi^{n/2-1}e^{-\psi}\,d\psi \\ &\propto \frac{1}{(ns^2+n(\bar{x}-\mu)^2)^{n/2}} \end{aligned}$$

where we use the fact that $\int_0^\infty f_{\mathrm{Gamma}(n/2,1)}(\psi)\,d\psi=1$. We lastly transform $t=\frac{\mu-\bar{x}}{S/\sqrt{n}}$, to transform the probability density function of $\mu|x$ into that of $\tau|x$, giving

$$\begin{aligned} f(\tau|x) &\propto f(\mu|x)\left|\frac{d\mu}{dt}\right| \\ &\propto \left(ns^2+n\frac{S^2t^2}{n}\right)^{n/2} \\ &\propto \left(1+\frac{t^2}{n-1}\right)^{-n/2} \end{aligned}$$

Note that $\left|\frac{d\mu}{dt}\right|$ is constant, so is absorbed by $\propto$. In the last line we use that $ns^2=(n-1)S^2$. By Lemma 1.2.5 we identify $\tau|x\sim$ Student-t$(n-1)$, as required.

7.7 We have

$$\begin{aligned} B &= \frac{\mathbb{P}[\Theta|_{\{X=x\}}\in\Pi_0]\mathbb{P}[\Theta\in H_1]}{\mathbb{P}[\Theta|_{\{X=x\}}\in\Pi_1]\mathbb{P}[\Theta\in H_0]} = \frac{\frac{1}{Z}\int_{\Pi_0} f_{M_\theta}(x)f_\Theta(\theta)\,d\theta\;\mathbb{P}[\Theta\in H_1]}{\frac{1}{Z}\int_{\Pi_1} f_{M_\theta}(x)f_\Theta(\theta)\;d\theta\mathbb{P}[\Theta\in H_0]} \\ &= \frac{\int_{\Pi_0} f_{M_\theta}(x)\frac{f_\Theta(\theta)}{\mathbb{P}[\Theta\in H_0]}\,d\theta}{\int_{\Pi_1} f_{M_\theta}(x)\frac{f_\Theta(\theta)}{\mathbb{P}[\Theta\in H_1]}\;d\theta} = \frac{\int_{\Pi_0} f_{M_\theta}(x)f_{\Theta|_{\{\Theta\in H_0\}}}(\theta)\,d\theta}{\int_{\Pi_1} f_{M_\theta}(x)f_{\Theta|_{\{\Theta\in H_1\}}}(\theta)\;d\theta} = \frac{f_{X|_{\{\Theta\in H_0\}}}(x)}{f_{X|_{\{\Theta\in H_1\}}}(x)}. \end{aligned}$$

Here, the third equality is a consequence of Theorem 3.1.2. The second-to-last inequality uses Exercise **1.8** and the final equality uses Exercise **3.6**.